internet retailer
STRATEGIES FOR MULTI-CHANNEL RETAILING

webselling views

Scientific Site Strategy

Navigating the depths of multivariable testing

By Gordon H. Bell and Roger Longbotham

Multivariable testing has made a big splash in the last few years with online retailers, yet this sudden success is really riding the crest of a wave that's been building for years. Multivariable testing—also called scientific testing, multivariate or matrix testing, Taguchi methods, or other branded terms—is based on a specialized field of statistics that has evolved over the last 80 years. Since the 1930s, a small group of academic statisticians has developed new test designs and techniques focused on efficient ways to test more variables more quickly.

Often called "experimental design," this specialty falls outside of mainstream statistics and has remained largely unknown to the business world. Only in the last decade have practitioners found a successful approach for using this impressive depth of academic theory to navigate fast-moving marketing channels.

Many variables at once

The concept is simple: with the right techniques you can change many variables at once—but in an organized way—so you can separate the impact of each. Complex mathematical principles define the "organized way" you need to set up your multivariable test.

The depth of statistical complexity below the surface can seem daunting. As marketers, you should understand the fundamental concepts and basic pros and cons of the selected test strategy. The expert who guides you through the process should be able to explain the rationale of his approach and have a good grasp of the vast realm of techniques available. These include efficient test designs like full-factorial, fractional-factorial and Plackett-Burman designs, plus a veritable A-to-Z of specialized tools: axial runs, Bonferroni method, confounding, dispersion effects and experimental units, plus orthogonality, projectivity and quadratic effects, down to the X-, Y- and Z-components of interaction.

Gordon H. Bell (above) is president of LucidView, a marketing consulting firm specializing in scientific testing techniques, and can be reached at gbell@lucidview.com. Roger Longbotham is senior statistician at Amazon.com Inc., where he oversees the multivariable tests on Amazon.com and conducts data mining studies related to customer behavior. He can be reached at longboth@amazon.com.

Various designs and techniques are appropriate for different marketing programs and objectives. For example, Plackett-Burman designs work well for testing 10-20 creative elements very efficiently in high-production-cost direct mail programs. Fractional-factorial designs are flexible and powerful for testing 5-15 creative elements and select interactions in e-mail and Internet programs. For product, price and offer testing—where elements are known to be important and interactions can be very large and valuable—full-factorial designs often are best. The number and type of test elements, cost and constraints on the number of "recipes" you can create, and the desired speed and precision of the test are among the issues that impact your choice of test design and strategy.

Since the dawn of direct marketing, split-run techniques (also called A/B splits, test-control or champion-challenger testing) have been the standard for marketing testing. You may have a long-running (or "control") banner advertisement and test it against one other with only the tagline changed, so any difference in click-through and conversion can be attributed to this one variable alone.

In contrast, one multivariable test design is made up of a number of related test "recipes." Instead of the one-variable change of a split-run test, one new banner ad in a multivariable test would include a number of changes—perhaps the new tagline along with a control graphic, new price, additional starburst and control background color. These multiple versions each has a unique combination of all elements in the test, each providing one new piece of data on every test element. Analyzing all recipes together, but grouping data in different ways, you can separate the

Multivariable vs. split-run

Benefit	Scientific multivariable testing	Split-run testing
Efficiency	2-35 variables in one test	1 variable=1 test
Speed and sample size	Constant sample size (no matter how many variables)	Increasing sample size
Depth of insights	Main effects, comparative effects, and interactions	One main effect (Cannot see interactions)
Flexibility	Wide range of test designs	One choice

precise impact of each change. The statistical structure requires that the creative execution accurately follows the defined test recipes.

Scientific multivariable tests have four key advantages over split-run techniques. You can test many marketing elements at once, using the same small sample size as A/B split, with results that quantify the impact of each element alone (main effect) and in combination with others (interaction), and with a vast array of techniques available to customize your approach.

The best e-mail recipe

A large Internet retailer/cataloger wanted to increase e-mail conversion. With 2-3 e-mail drops per week to a customer base of 450,000, conversion rate averaged 1% per campaign. The team had a challenge pinpointing what worked best because they continually changed e-mail creatives and offers to keep the program fresh.

After brainstorming 42 ideas, the team narrowed the list down to 18 bold, independent test elements for one multivariable test made up of 20 different combinations, or recipes, of all 18 elements. Four of these "recipes" are shown on p. 75 (with control levels in black and the new ideas in orange). A direct subject line might be something like "Save 20% through Friday." A creative subject line would change that to "Awesome new products and super savings that won't last."

Recipe 20 was simply the control. All other recipes had about half the elements set at the control level and half at the new level, but a different half-and-half for each recipe. Though these four may look like random combinations, all recipes fit within the precise statistical test design. Like the pieces of a puzzle, all recipes fit together to provide accurate data on the main effects and important interactions of all 18 elements.

The team ran the same test across three different types of promotions to see promotion-specific effects plus elements that were important across all campaigns. Results for the first campaign are shown below, including the 18 main effects (shown in the bar chart) and one key interaction (in the line plot).

In the chart, main effects are arranged from the largest (D, at the top) to the smallest. Test elements are listed on the left with the "new idea" shown in parentheses. The length of the bar and the label show the size of the main effect. The +/- sign on the effect shows whether the new idea is better (positive effect) or the control is better (negative effect). The dashed line of significance is a measure of experimental error. All effects below that line (less than 6%) can be explained simply by random variation. Effects are shown as a percentage change from the control, so a 10% effect would increase conversion rate from 1% for the control to 1.1%.

Four main effects were clearly significant: Product selection had the largest effect. Conversion rate increased by 10% when best-selling products (D+) were promoted instead of "unique" products. A larger headline with color (J+) increased conversion by 8.9%. Offering three products decreased conversion by 8.2% vs. one product (E-). Finally, the creative subject line (A+) beat the direct subject line by 6.9%.

Immersed deeper within the unique statistical structure is a wealth of information about interactions. On the surface, main effects show individual changes that increase conversion. Interactions show how these effects may ebb or flow depending on the relationship among marketing-mix elements.

The line plot on p. 76 shows the AB interaction. The main effect of A (subject line theme) changes significantly depending on whether certain words are capitalized in the subject line (element B). Supporting the main effect in the bar chart, the creative subject line is always better than the direct offer (going from left to right), but the impact is much greater with no capitalization (B+, orange line). This interaction shows that (1) capitalizing words in the subject line does have an impact (B+: no capitalization) and that (2) without capitalization, the effect of the creative subject line (A+B+) is about 40% larger than shown by the main effect in the bar chart. Interactions not only offer deeper insights into the true relationship among elements, they also help better quantify the impact of the optimal combination of elements.

What's the difference?

In this case, what was the advantage of using multivariable testing? Well, if the team had used simple split-run techniques instead:

Finding the right e-mail creative

Samples of e-mail marketing campaign tested recipes (new ideas in orange, control elements in black)

Test Elements	Recipe 1	Recipe 2	Recipe 3	Recipe 20
A. Subject line theme	Direct	Creative	Direct	Direct
B .Capitalization in subject line	Yes	Yes	No	Yes
C. Spam filter friendly subject	Yes	Control	Control	Control
D. Product selection	Unique selection	Best sellers	Best sellers	Unique selection
E. No. of products	1	3	3	1
F. Product image size	Control	Large	Large	Control
G. Photo background	Plain	Environmental	Plain	Plain
H. Copy style	Copywriter B	Copywriter A	Copywriter B	Copywriter A
I. Copy length	Long	Long	Long	Short
J. Headline style	Large with color	Large with color	Control	Control
K. Background color	White	White	Color	White
L. Starburst	Yes	Yes	No	No
M. Banner at top	Yes	No	Yes	No
N. Button text	Buy now	Click here	Buy now	Click here
O. Button graphic	3D	3D	3D	Control
P. No. of links	Control	Control	Control	Control
Q. Below-the-fold content	None	None	None	None
R. Expiration date	Short	Control	Control	Control

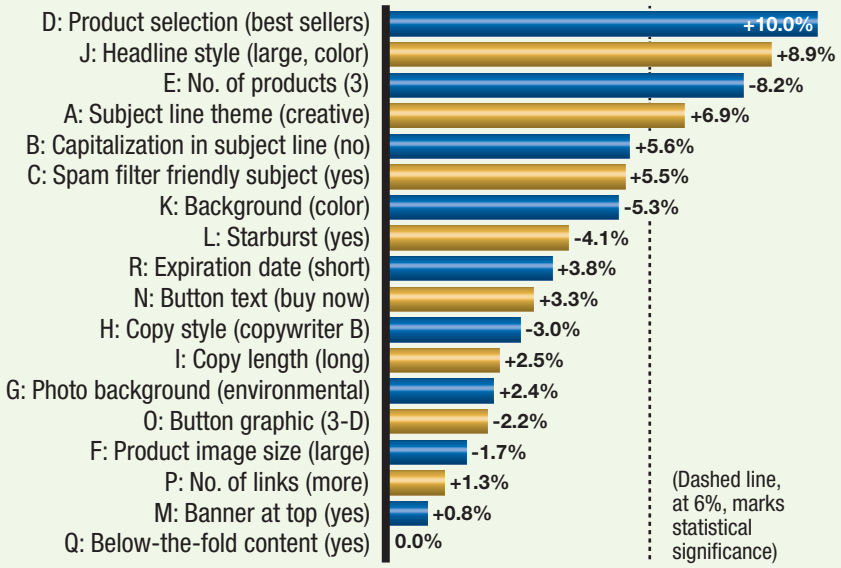
- Testing all 18 elements in one drop, not one effect would have been significant, since the line of significance would have been 2 1/2 times higher (only effects greater than 16% would be significant).
- For equal confidence, the team would have to test only one element per drop, requiring 18 campaigns, with no way to separate seasonality or differences among campaigns.
- The team would never have seen the AB interaction (and others) and capitalization would appear to have no impact.

An extreme case of multivariable testing was one banner ad test of 26 elements. Testing 10 graphical elements, 9 messages, pop-ups, drop-downs, animation and other marketing tactics, the biggest challenge in this test was defining the elements and managing recipes to avoid completely absurd combinations (imagine all these words and graphics stuffed into one banner). In cases like this, the "art" of testing—defining clear, bold, independent test elements and creating a test design with recipes that push the limits of market knowledge without falling apart in execution—is equally as important as the science. In this case, eight significant main effects and one very profitable interaction led to a 72% jump in conversion. The test was completed in four weeks. For equal confidence, split-run tests would have required 14 months.

An Internet retailer of consumer gifts ran a landing page test of 23 elements for three weeks and pinpointed seven changes to increase sales (and six "good" ideas that hurt) for a 14.3% jump in sales. 23 separate A/B splits would have required over 40 weeks to achieve equal statistical confidence. This test paid for itself 10 days after results were implemented.

Main effects: conversion rate (campaign 1)

(Effects as % change from the control)

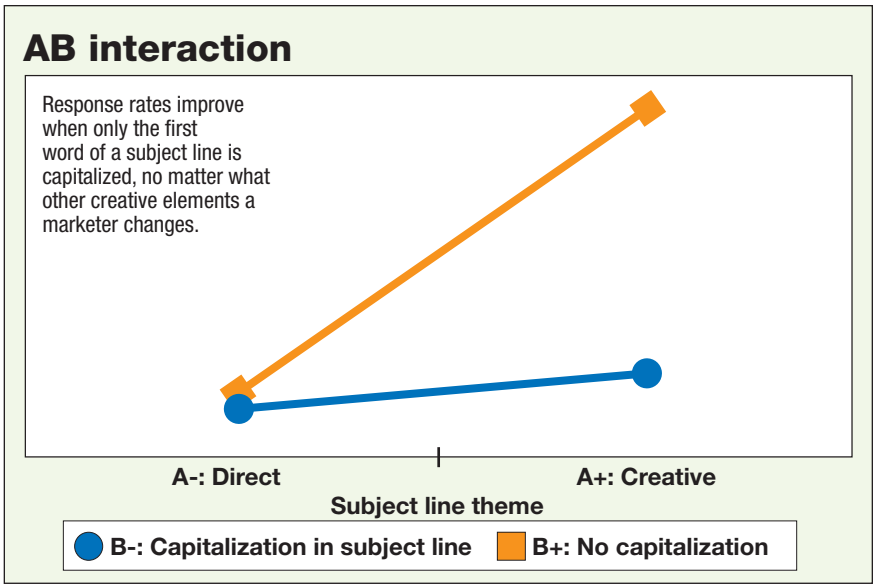


Getting started

Multivariable testing is most effective for retailers who have many ideas to test and the flexibility to create numerous recipes within a high-value marketing program. Key decisions in launching a retail test include: choosing the right

test elements and levels for each, deciding which of all possible combinations should be executed for a valid test, creating and executing all recipes, and collecting data and analyzing results.

Software platforms offered by firms like Optimost and Offermatica can simplify the process of creating recipes and analyzing results for Internet tests. Consultants focused on the specialized statistics and strategies of testing can help guide you through the process, especially for e-mail and offline programs (like direct mail, media advertising and in-store tests). For outside assistance, you may want to budget about $10,000 per month for ongoing support. In return you can expect to reduce your learning curve and increase your testing efficiency and return on investment. Another option for small firms is the free, barebones service from Google for Adwords advertisers called the Website Optimizer.



Testing remains an integral part of every good marketing program. Like trading in your dinghy for a clipper ship, launching a multivariable test brings you the power and freedom to move faster through turbulent marketing channels. With an experienced guide to show you the way, scientific testing offers greater agility to respond to market changes, streamline your retail programs and explore new opportunities for growth. ●