The Basic Science behind Multivariable Testing – A Search Ad Case Study

by Gordon H. Bell
published on AudienceDevelopment.com
September 2 and 9, 2008

Scientific multivariable testing is gaining traction in the marketing world because it offers deeper insights with greater accuracy, smaller sample sizes, and more efficient use of resources than common split-run techniques. One big challenge is that multivariable testing is not one thing, but a collection of diverse (and sometimes complex) statistical techniques and test designs developed over the last 80 years. With 700-page academic tomes and a large gap between theory and application, gaining the statistical expertise may be seem a daunting task. However, like the performance of a virtuoso violinist, a well-executed test should be very accessible and easy to appreciate even if you're not an expert yourself.

The following case study walks through the development, execution, and analysis of a basic multivariable test. Once you understand the basics, most other test designs are variations of the same theme.

Case study: search ad test

Using a LucidView example, we can present all the details of the test. This project had a simple goal: increase the impact of the paid search ads placed on the Google network. With four lines in each ad, we defined four test elements:

	Test Elements	(-) Control	(+) New idea
A	Title	Successful Internet Tests	e-mail, e-tail, e-Testing
B	Line 1	with the power and precision of	Drive online ROI with guidance in
C	Line 2 with "recipe"	scientific multivariable testing	multivariable testing (recipe #X)
D	Destination page	www.lucidview.com/Internet_testing	www.lucidview.com

Element A tests two options for the ad title and element B tests the first line below the title. Element C tests the addition of "(recipe #8)" (or another number) at the end of line 2 as a hint that the ad itself is part of a test. The destination page (element D) is not displayed, but sends the visitor either to the home page (D+) or an Internet-specific webpage (D-).

The multivariable test design – like threads in layers of fabric

These four ideas can easily be tested as four separate A/B splits. But the benefits of rolling them into one multivariable test far outweigh the effort (as we'll see next week in part 2 of this case study).

With split-run testing, each variable is a separate statistical test. With advanced techniques, many variables can be tested within the same test design. How? By weaving the A/B tests together like different color threads in a piece of fabric. A statistically-valid scientific test creates a pattern of different layers of cloth (test cells or "recipes") from numerous colors of

thread (marketing-mix variables). In this dense fabric, each thread remains distinct, but the combination becomes stronger and more stable. The unique pattern offers deeper insights and greater statistical power "per square inch" (i.e., for the given sample size).

Here's how it works…

You can weave a simple multivariable test from two "threads." Starting with elements A and B (above), we can layout a test with four "layers," as follows:

This matrix uses "–" to represent the control setting, above, and "+" to represent the new idea, or test setting. Each column is a different thread in our test fabric and each row is a different layer (and a different version of the ad).

	Title	Line 1	
Recipe	A	B	
1	–	–	Control ad
2	+	–	New title
3	–	+	New line 1
4	+	+	Both changed

The first three recipes are exactly like split-run tests: the control (recipe 1) versus the new title (recipe 2) and the new line 1 in the ad (recipe 3). Adding recipe 4 makes this a balanced scientific test. Here's why…

Of these four ad versions, two have the new title (recipe 2 and 4) and two have the control title (recipe 1 and 3). Of the two A+ recipes, one has B- (control line 1) and one has B+ (the new line 1). Here's the clincher: if you *average* recipes 2 and 4, you get the average impact of the new title, but the impact of changing line 1 simply averages out. No matter how large or small the difference between B- and B+, by averaging an equal number of both, the effect of B cancels out.

By adding one more recipe (the all-plus recipe), we create this perfect pattern that allows us to test two variables at once and still separate out the impact of each. Similar to an A/B split, the effect of A is simply the average performance of ads 2+4 minus the average of ads 1+3. The effect of B is the average of recipes 3+4 versus the average of recipes 1+2.

Scientific tests offer clear advantages

A balanced multivariable test allows us to use *all* the data to analyze each test element, so unlike A/B splits, we *do not* need additional sample size for each new variable! We can use the same sample size whether testing 2 or 22 elements at once – as long as all the elements are woven together in the same test design.

Scientific tests also allow us to analyze interactions between test elements. In this case, we can clearly analyze if changing the title (A+ versus A-) has a different impact with the control line 1 than with the new line 1 (by analyzing the difference between recipes 2 versus 1 and recipes 4 versus 3). Interactions may lead to a different rollout decision and are especially valuable for price and offer testing.

4-element Search Ad Test

This same concept – weaving together many test elements into layers within the same fabric – can be extended to larger, more complex test designs. The LucidView paid search test included all four elements tested within 8 different versions of the ad, following the test design below.

The balanced pattern extends to this larger test: averaging all A+ recipes (2, 4, 6, 8), each other column has two pluses and two minuses, so every other element cancels out. The same is true for the four A- recipes, as well as both levels of B, C, and D. This perfect balance is difficult to achieve, yet straightforward to understand.

	Title	Line 1	Line 2 with "recipe"	Destination page
Recipe	A	B	C	D
1	–	–	–	–
2	+	–	–	+
3	–	+	–	+
4	+	+	–	–
5	–	–	+	+
6	+	–	+	–
7	–	+	+	–
8	+	+	+	+

Test Execution

All versions of the paid search ad were created following these eight recipes and placed in Google AdWords to run at random using the same "Internet testing" related keywords and CPC.

The control ad (recipe 1) and "all plus" ad (recipe 8) are shown below. The other six ads are different combinations of these lines of text (and destination page) following recipes 2-7.

Results

After a number of weeks, we collected and analyzed click-through data, shown below. (Certainly other metrics – like visit length, requests for information, and new accounts – are more valuable than the click-through rate (CTR), but those metrics are not presented here.)

Looking at the click-through rate (CTR) of all 8 recipes, you see that recipe 6 is the winner. However, this ignores information from all the other recipes. As discussed earlier, a multivariable test is analyzed column-by-column, using data from all 8 recipes grouped in different ways (depending on where the minuses and pluses fall). After each effect is calculated, the winning ad is *created* from the optimal combination of all significant effects – and the winner may not even be one of the ads tested!

	Title	Line 1	Line 2 with "recipe"	Destination page			
Recipe	A	B	C	D	Impressions	Clicks	CTR
1	–	–	–	–	3843	16	0.42%
2	+	–	–	+	5311	51	0.96%
3	–	+	–	+	3856	15	0.39%
4	+	+	–	–	5214	53	1.02%
5	–	–	+	+	4312	16	0.37%
6	+	–	+	–	5333	70	1.31%
7	–	+	+	–	3819	10	0.26%
8	+	+	+	+	5189	67	1.29%
Main effects:	*0.785%*	*-0.025%*	*0.110%*	*0.000%*			

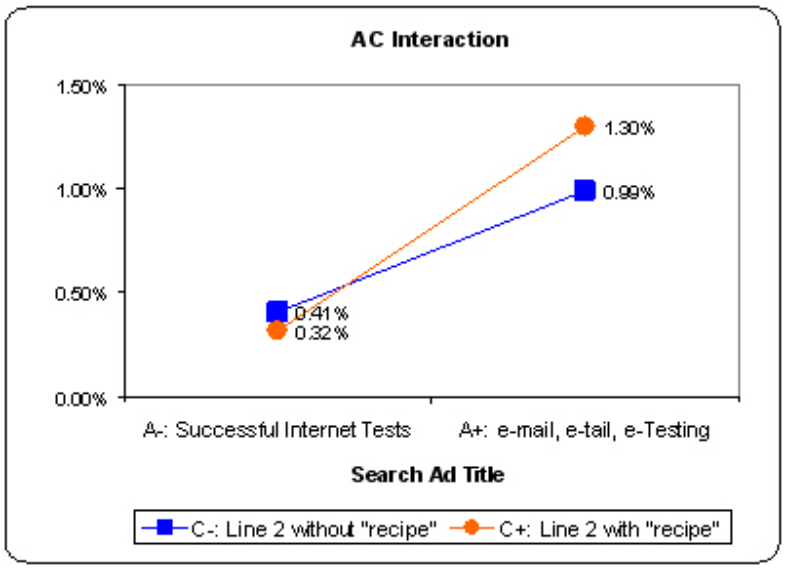
Calculating main effects

The main effect of A is calculated as the average CTR of recipes 2, 4, 6, and 8 minus the average of recipes 1, 3, 5, and 7. This means that the new title, "e-mail, e-tail, e-Testing" increases CTR by 0.785% over the control title, "Successful Internet Tests" (nearly a 3x increase versus the control of 0.42%). The main effect of B is calculated as the average of all B+ recipes (3, 4, 7, 8) minus the average of all B- recipes (1, 2, 5, 6). The main effects of C and D are calculated the same way.

The next step is calculating the Line of Significance showing how large an effect must be to rise above the noise. In this case, you can rearrange the sample size equation (shown in this article [http://www.circman.com/viewMedia.asp?prmMID=3676]) using the total number of impressions (36,877), the average click-through rate (0.81%), and the constant 15.37 (in place of 31.38) to calculate the "smallest lift" that is statistically significant as ±0.18%. Therefore, only the main effect of A is significant.

Interactions offer deeper insights

But wait – we can also analyze interactions to look for deeper insights among these elements. Using this test design, interactions are calculated much the same way as main effects, but looking at pairs of columns together. The one significant interaction effect that stands out is the AC interaction, pictured below.



The interaction plot looks at how the main effect of A (moving left-to-right) changes depending upon the setting of C (orange line for C+ versus the blue line for C-) and vice versa. You see that with the control title, there is not much difference between C+ and C- (the blue and orange points at the lower left). But with the new title, C+ (line 2 with "recipe") does have a positive impact (orange point in upper right).

Although the average impact of C is insignificant, the two-way interaction uncovers a new opportunity. With the new title, changing line 2 to "multivariable testing (recipe #X)" increases CTR to 1.3%.

With these results, the optimal search ad is:

e-mail, e-tail, e-Testing
with the power and precision of
multivariable testing (recipe #6)
www.LucidView.com

How does multivariable compare with split-run testing?

Instead of this fairly simple multivariable design, if we had tested the same ideas as A/B splits:
1. We could have used only 5 ads (control + 4 test cells) instead of 8, but…
2. The AC interaction would be impossible to see
3. Experimental error would have been 2x greater
4. The test would have to run 4x longer for equal statistical confidence

Scientific multivariable testing is based on complex mathematics, but the process can still be straightforward. A skilled craftsman can help you create an intricate yet robust fabric, weaving many marketing-mix variables into one clear, actionable, multi-layered test design. With hundreds of possible test designs and techniques, you can select just the right design to match your objectives. If testing is an important part of your marketing programs, multi-layered scientific testing offers greater power and freedom to see your marketplace more clearly.

Gordon H. Bell is president of LucidView, a consulting firm helping industry leaders increase marketing ROI and learn best practices in the art and science of marketing testing. He can be reached at www.lucidview.com or 888-lucidview.