

Robust Test Designs for Dynamic Marketing, Retail, and Advertising Programs

Bell, Gordon
LucidView
80 Rolling Links Blvd.
Oak Ridge, TN 37830 USA
E-mail: gbell@lucidview.com

Introduction

Marketing and advertising managers began “scientific advertising” in the early 1900s with split-run newspaper tests. Large publishers printed two runs of the same newspaper. Placing different versions of their advertisement (with a unique mail-in coupon) in each run, managers could measure the number of orders from each. This offered a simple way to test new pictures, copy, prices, and other variables, simply by making one change to the ad in the second run. For the last 90 years, the majority of marketers and advertisers have continued to use the same one-variable-at-a-time testing method.

The birth of design of experiments began around the same time. But as this specialized field grew, the marketing world persisted with one-variable tests, largely unaware of multivariable methods. The chasm between the science and practice of marketing testing continued. Even today, most market leaders have at best a passing awareness of scientific “multivariable testing” (as design of experiments has come to be called in marketing).

Challenges of In-Market Testing

The unique challenges of testing in marketing, retail, media advertising, and Internet programs also create immense opportunity for applying advanced experimental design techniques (this paper uses “test” and “experiment” interchangeably). The primary challenge is the lack of market control. Controlled environments—whether engineering laboratory, manufacturing plant, or even market research study—have seen more widespread use of experimental design. Closer to the “front lines” of marketing, statistical issues are overshadowed by myriad human, operational, cost and control issues. The textbook statistics are secondary to “getting the mail out the door... closing the sale... staying on budget...” and managing the daily demands of every marketing program. Like the framing of a house, the statistical structure must be built correctly on a solid technical foundation. Yet all the elements of the marketing-mix built upon that framework ultimately determine the attractiveness and value of the outcome. Successful marketing testing combines the efficient use of powerful experimental design techniques along with skill in the art of marketing and test execution. In spite of the challenges, large marketing investments and a high potential financial return are making experimental design an important tool for improving performance in competitive markets.

Unique Requirements of Experimental Design in Marketing

Factorial, fractional-factorial, and other experimental designs have been widely used in manufacturing, but seldom in marketing, retail, and advertising applications. The few published examples tend to focus on market research applications and small retail tests using full-factorial or fractional-factorial experimental designs.

Recently, market leaders have achieved considerable benefit from large in-market fractional-factorial (and related) test designs¹. This may be due, in part, to the unique characteristics of marketing testing not found in a manufacturing environment:

1. A large number of potential test factors
2. Few well-established “laws” of marketing
3. Constant change in the marketplace (high variability and instability)
4. Limited window of opportunity – both for testing and applying results

A nearly endless list of marketing-mix variables can be tested. In particular, creative changes—like the words, pictures, and layout of a direct mail package—can be defined and tested in numerous ways. Whereas a manufacturing process often has a discrete number of “knobs” that can be adjusted, the intangible nature of marketing programs removes many physical limits on the potential number of factors. For example, even a simple concept like “price” is very flexible. The absolute price (e.g., \$9.99) cannot be separated from the creative presentation: the number of times price is shown, along with the size and location, highlights and starbursts, and other ways of presenting that price point to the consumer.

Physical properties of materials and manufacturing processes are often well understood. “Laws” of the marketplace are more nebulous. The impact of the words in a newspaper ad, or color of an envelope, or product packaging on supermarket shelves cannot be distilled down to a physical law or mathematical equation. Expert copywriters, graphic artists, and retailers do follow established rules of effective marketing, but these rules encompass greater variability and subjectivity than the laws of nature. There are simply many questions that one can ask in marketing—numerous ideas than can become valid test factors.

Experimenting on an unstable process is difficult, yet markets tend to be in constant flux. Customers, products, marketing programs, and the competitive environment learn, evolve, change, and react. Instability may be an objective in itself—catch the customer’s attention with an exciting new product and offer, presented in a unique mailing, e-mail, or advertisement that stands out above the “common cause” competition. In effect, marketers regularly try to create a “special cause.”

For this reason, factors are best tested at a stable point within an unstable market. Testing many variables at once—in one mailing, randomized across one group of customers, using the same products and promotions—can provide statistically valid results. Though results cannot necessarily be extrapolated to future campaigns, proving the significance and hierarchy of effects and interactions at a single point in time gives marketers clear insights to help them plan future campaigns. With the speed of change, immense market uncertainty, and endless possibilities, large fractional-factorial tests provide an immense quantity of valuable insights.

In dynamic and “noisy” markets, speed and test power are imperative. In-market tests often require significantly larger sample sizes than manufacturing experiments, not only due to market noise, but also because many tests focus on response rate (binomial versus variables data)—does someone respond to the mailing or e-mail, or buy a product online or in the retail store. With the normal response to a credit card offer perhaps 0.4%, a test mailing of 500,000 is common. The speed of market change also leads to a focus on short-term metrics and rapid implementation of results.

The variability of factor effects on purchase behavior tends to increase as testing moves “upstream” from the point of purchase. Therefore, in-market testing can measure buyer behavior better than market research studies measuring opinion or intent. Even within in-market tests, variability increases as the test is executed farther back from the point of purchase. For example, in the retail environment, in-store tests tend to have less noise than advertising tests. A newspaper advertisement or circular reaches a broader audience than the population of potential buyers and is separated by some length of time and distance from the purchase decision. Each step in the process—between exposure to the advertisement and payment at the cash register—adds a layer of variation to the test. Media advertising tests can be valuable, but the sample size (number of people exposed and/or the length of time the test runs) must often be greater. This is likely one reason why experimental design appears more widely used in marketing channels with a direct-response mechanism (like direct mail, catalogs, e-mail and Internet campaigns). That said, the financial investment and potential return in retail and advertising testing is so large that these applications should be pursued, simply with larger sample sizes. In addition, the selection of bold factors and stable test units can help overcome the challenges of high variability.

Fractional-factorial Designs

One downside of experimental design in marketing is that each treatment combination (also called a test recipe) can be costly to develop and execute. Website and e-mail tests can have lower per-recipe costs and therefore use larger test designs. In contrast, direct mail and advertising tests often require efficient designs with minimal test recipes. The selection of test design is primarily driven by:

1. The number of factors desired
2. Interest in analyzing interactions along with main effects
3. The cost of additional test recipes
4. The number of levels tested for each factor

If test recipes were free, every test could use a full-factorial design. This would provide an independent calculation of every main effect and all possible interactions. In the following direct mail case study (starting on page 5) with 19 factors, each at two levels, a full-factorial design would include 524,288 versions of the mailing (2^{19} test recipes). Clearly this is not possible, so some sort of fractional (or non-orthogonal) test design is necessary. Effect sparsity, effect heredity, and effect dissipation (the author's term leveraging the earlier concept of hierarchical ordering) justify fractional-factorial designs with the principles that:

- Few effects tend to be significant
- Interactions tend to include factors with large main effects
- Two-factor interactions tend to be smaller than main effects and higher-order interactions (3-factor, 4-factor, and so on) rapidly disappear.

More factors require a larger minimum test design. Seven two-level factors requires at least eight test recipes; 15 factors requires at least 16 test recipes. Small test designs are appropriate when the marketing team wants to focus on a few variables already proven important. Large test designs are an efficient way to test many new ideas at once to rapidly optimize a marketing program by identifying the most important tactics (and avoiding costly changes that have little effect).

Designs with a lower ratio of recipes-to-factors primarily test main effects. If interactions are important or may be large (as is often the case with price testing), then a test of more combinations of fewer variables may be most appropriate. However, "low resolution" main-effects designs (in which many two-factor and higher-order interactions are correlated with main effects) can allow for a limited analysis of select two-factor interactions, as shown in the case study, below.

Higher per-recipe costs usually dictate a smaller test design. A 32-recipe fractional-factorial design may be most effective for a 15-factor Internet test (to reduce "confounding error" at minimal cost), but a 16-recipe design may be preferable for a direct mail test.

Orthogonal versus Optimal Test Designs²

In marketing tests, the number of levels for each factor is nearly as limitless as the potential list of factors. Unfortunately, a test of three or more levels can lead to a factorial or fractional design with an unmanageable number of recipes. Design of experiments offers one solution with optimal designs. Whereas orthogonal factorial and fractional designs have zero correlation among columns in the test matrix (all main effects and, at times, certain interactions can be estimated independently), optimal test designs have some correlation among effects but are designed to limit the number of recipes while optimizing a selected design characteristic. For example, D-optimal designs are set up to minimize the volume of the confidence region of the effect estimates (considering the variances and covariances of these estimates).

Optimal designs allow you to test under sub-optimal conditions where certain combinations are constrained, the cost of testing is immense, or the "response surface" has abnormal characteristics. Fortunately, these constraints are seldom necessary in marketing tests. In marketing testing, the most "optimal" test design is usually one with a straightforward execution, clear analysis, and easily-understood results.

Instead of testing multiple levels at once, marketers may decide to:

- Test bold levels in an efficient two-level design, followed by further “refining” tests, analyzing significant effects at new levels (running a series of tests similar to response surface methodology).
- Break multiple levels into multiple test designs. For example, if two customer segments cannot receive the same offer terms, each can be tested in a separate orthogonal design, rather than combining both into one optimal design with a constrained design region.
- Use centerpoints to test for curvature of the effects—adding one test recipe with all factors set at a level directly between the high (+) and low (-) levels. If curvature is significant, then further tests (or additional design points) can be run to further assess curvature.
- Test three or four levels along with a 2-level orthogonal test design. A three- or four-level factor can be included within a two-level test matrix, but this adds complexity. Often a more efficient strategy is to add one or two additional “splits” (one-variable test recipes) onto a standard 2-level test design, allowing for the analysis of additional levels of one factor (but requiring additional sample size) without detracting from the robust two-level design.
- Define multi-level ideas as a combination of two-level factors.

For example, a Conde Nast Publications e-mail test³ included four ways to present the special subscription price: with a simple message, adding “that’s less than 85¢ an issue” highlighted in a red dot, adding “that’s like getting 18 issues free” in the text, and showing a graphic with the cover price crossed out and the special subscription price highlighted. Instead of defining a 4-level factor, this “price presentation” concept was defined as three separate factors:

	<u>(-) level</u>	<u>(+) level</u>
"85¢ an issue" dot	No	Yes
"Like 18 issues free" text	No	Yes
Strikethrough pricing graphic	No	Yes

The test design included all combinations of these three factors among the test recipes: recipes with all three “no” (a simple price message without any of these) and recipes with one, two, and all three of the factors. Not only did this eliminate the need for a 4-level factor, but also provided data on interactions among all three.

Plackett-Burman Designs

Fractional-factorial techniques offer a way to test a large number of factors with a fraction of the test recipes. One practical drawback is that all fractional designs are a power of two, with no alternatives between designs with 4, 8, 16, and 32 (and so on) test recipes. 8-15 factors require a 16-recipe fractional design. Adding one more factor then doubles the size of the design—testing 16 factors in 32 recipes.

Plackett-Burman designs are orthogonal two-level designs similar to fractional designs, but with the number of recipes a multiple of four. Plackett-Burman designs with 12, 20, 24, and 28 recipes (and so on) offer alternatives to minimize the number of test recipes.

The confounding schemes of Plackett-Burman designs are different than with standard fractional designs. In a fractional-factorial design, interactions are fully-confounded in one column of the test matrix. In contrast, Plackett-Burman designs have interactions partially confounded in a number of columns. Two benefits are that (1) each interaction does not add as much “confounding error” to any one effect and (2) certain interactions may be estimated along with main effects, whereas in a fractional design, interactions and main effects have 1:1 correlations, so cannot be

separated mathematically. Two drawbacks of Plackett-Burman designs are (1) the complex confounding scheme and (2) that each interaction—though not fully confounded with any one column—is partially confounded within numerous columns.

The complexity of Plackett-Burman designs is likely the reason for their limited application historically. However, as the following case study shows, a well-run Plackett-Burman design can overcome these drawbacks and offers a method for testing more factors very efficiently.

Direct Mail Test: 19 Factors in a 20-recipe Plackett-Burman Design⁴

The financial industry—including insurance, investment, credit card, and banking firms—was among the first to use experimental design techniques for marketing testing. The project described here is from a leading Fortune 500 financial products and services firm. The company name and proprietary details have been removed, but the test strategy, designs, results, and insights are accurate.

The firm’s marketing group regularly mailed out credit card offers and wanted to find new ways of increasing the effectiveness of their direct mail program. The 19 factors shown in the table, below, were thought to influence a customer’s decision to sign up for the advertised product. Factors A – E were approaches aimed at getting more people to look inside the envelope, while the remaining factors related to the contents and offer inside. Factor G (sticker) refers to the peel-off sticker at the top of the letter to be applied by the customer to the order form. Factor N (product selection) refers to the number of different credit card images that a customer could chose from, while the term “buckslip” (factors Q and R) describes a small separate sheet of paper that highlights product information.

The 19 test factors and their low and high levels (called the “control” and “new idea”)

Factor	(-) Control	(+) New Idea
A Envelope teaser	General offer	Product-specific offer
B Return address	Blind	Add company name
C "Official" ink-stamp on envelope	Yes	No
D Postage	Pre-printed	Stamp
E Additional graphic on envelope	Yes	No
F Price graphic on letter	Small	Large
G Sticker	Yes	No
H Personalize letter copy	No	Yes
I Copy message	Targeted	Generic
J Letter headline	Headline 1	Headline 2
K List of benefits	Standard layout	Creative layout
L Postscript on letter	Control version	New P.S.
M Signature	Manager	Senior executive
N Product selection	Many	Few
O Value of free gift	High	Low
P Reply envelope	Control	New style
Q Information on buckslip	Product info	Free gift info
R 2nd buckslip	No	Yes
S Interest rate	Low	High

The 20-run Plackett-Burman main effects design, below, was created to test these 19 factors in the fewest test recipes.

The 20-recipe Plackett-Burman design with response rates for each mailing

Test Cell																Orders	Response Rate				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O			P	Q	R	S
1	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	-	52	1.04%
2	-	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	38	0.76%
3	+	-	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	42	0.84%
4	+	+	-	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	+	134	2.68%
5	-	+	+	+	+	+	-	+	+	+	+	-	+	-	+	-	-	-	-	104	2.08%
6	-	-	+	+	-	+	+	-	-	+	+	+	+	-	+	-	+	-	-	60	1.20%
7	-	-	-	+	+	-	+	+	-	+	+	+	+	-	+	-	+	-	-	61	1.22%
8	-	-	-	-	+	+	-	+	+	-	+	+	+	-	+	-	+	-	+	68	1.36%
9	+	-	-	-	-	+	+	-	+	+	-	+	+	+	+	+	-	+	-	57	1.14%
10	-	+	-	-	-	-	+	+	-	+	+	-	-	+	+	+	+	-	+	30	0.60%
11	+	-	+	-	-	-	-	+	+	-	+	+	-	-	+	+	+	+	-	108	2.16%
12	-	+	-	+	-	-	-	+	+	-	+	+	-	-	+	+	+	+	+	39	0.78%
13	+	-	+	-	+	-	-	-	-	+	+	-	+	+	-	-	+	+	+	40	0.80%
14	+	+	-	+	-	+	-	-	-	-	+	+	-	+	+	-	-	+	+	49	0.98%
15	+	+	+	-	+	-	+	-	-	-	+	+	-	+	+	-	-	-	+	37	0.74%
16	+	+	+	+	-	+	-	+	-	-	-	+	+	-	+	+	-	-	-	99	1.98%
17	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	-	+	+	-	86	1.72%
18	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	-	+	+	43	0.86%
19	+	-	-	+	+	+	-	+	-	+	-	-	-	-	-	+	+	+	+	47	0.94%
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	104	2.08%

In a Plackett-Burman design each pair of factors (columns) is orthogonal, which by definition means that each of the four factor-level combinations, (- -), (- +), (+ -), (+ +), appears in the same number of recipes. In the 20-recipe design (above) for every pair of columns, each of the four combinations appears five times. As a consequence of orthogonality, the main effect of one factor can be calculated independently of the main effect of all others. Plackett and Burman showed that the complete design can be generated from the first row of +’s and -’s. In this matrix, the last entry in row 1 (-) is placed in the first position of row 2. The other entries in row 1 fill in the remainder of row 2, by each moving one position to the right. The third row is generated from the second row using the same method, and the process continues until the next to the last row is filled in. A row of -’s is then added to complete the design.

Sample Size

The focus of the direct mail test was on increasing response rate: the fraction of people who respond to the offer. The overall sample size (the number of people to receive test mailings) was determined according to statistical and marketing considerations. The chief marketing executive wanted to limit the number of names to minimize the cost of test mailings performing worse than the control (especially when testing a higher interest rate) and to reduce postage costs. Of the 500,000 total packages that were mailed, 400,000 names received the “control” mailing (the current highest-response version of the mailing) that was run in parallel to the test, while 100,000 were used for the test. Therefore, each of the 20 test cells in Table 2 was sent to 5,000 people, resulting in the response rates listed in the last column.

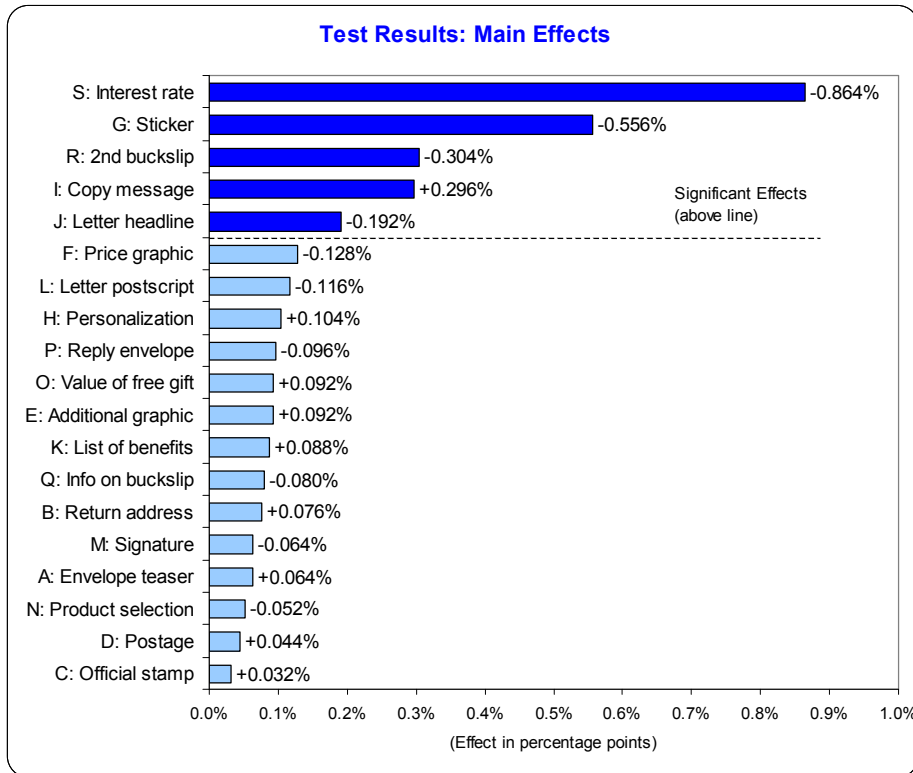
For each factor in the experiment, 50,000 people received a mailing with the factor at the plus level and 50,000 people received a mailing with the factor at the minus level. Each main effect is obtained by comparing average responses from these two independent samples of 50,000 each. Because the design is orthogonal, the same 100,000 people are used to obtain independent estimates

of each main effect. Using one-variable testing techniques, a sample size of about 25,000 would be required for each of 19 one-factor tests, for a total sample size of 475,000. Therefore, 375,000 more people would have to be tested to obtain the same statistical power as the Plackett-Burman design.

Test Results

The estimated effects, which are differences between average responses at the plus and minus levels of the factor columns, are shown below. In the chart, effects are ordered from the largest (at top) to the smallest (at bottom), in terms of their absolute values. The sign of each effect shows which level is better: For positive effects, the “+” level increases response; for negative effects, the “-” level increases response.

Main effects estimates: Plackett-Burman design



Significance of the effects was determined by comparing the estimated effects with their standard errors. The result of each experimental run is the proportion of customers who respond to the offer. Each proportion is an average of $n = 5,000$ individual binary responses; its standard deviation is given by $\sigma = \sqrt{\pi(1-\pi)/n}$, where π is the underlying true proportion. Each estimated effect is the difference of two averages of $N/2 = 10$ such proportions. Hence its standard deviation is

$$\text{StdDev}(\text{Effect}) = \sqrt{\frac{2}{N} \frac{\pi(1-\pi)}{n} + \frac{2}{N} \frac{\pi(1-\pi)}{n}} = \sqrt{4/N} \sqrt{\frac{\pi(1-\pi)}{n}}$$

Replacing the unknown proportion π by the overall success proportion (averaged over all runs and samples), $\bar{p} = (\# \text{Purchases}) / (nN) = 1,298 / 100,000 = 0.01298$, leads to the standard error of an estimated effect,

$$\text{StdError}(\text{Effect}) = \sqrt{4/20} \sqrt{\frac{(0.01298)(0.98702)}{5,000}} = 0.00072$$

The standard error is 0.072 if effects are expressed in percentage terms. Significance (at the 5 percent level) is determined by comparing the estimated effect with 1.96 times its standard error, $\pm 1.96(0.072) = \pm 0.141$. The dashed line in the chart separates significant and insignificant effects.

The following five factors had a significant effect on the response rate:

S- or Low interest rate: Increasing the credit card interest rate reduces the response by 0.864 percentage points. The financial gain from the higher rate would be much less than the loss due to the decrease in the number of customers.

G- or Sticker: The sticker (G-) increases the response by 0.556 percentage points, resulting in a gain much greater than the cost of the sticker.

R- or No 2nd bucksliip: Adding another bucksliip reduces the number of buyers by 0.304 percentage points. One explanation offered for this surprising result was that the bucksliip added unnecessary information.

I+ or Generic copy message: The targeted message (I-) emphasized that a person could chose a credit card design that reflected his or her interests, while the generic message (I+) focused on the value of the offer. The creative team was certain that appealing to a person's interests would increase the response, but they were wrong. The generic message increased response by 0.296 percentage points.

J- or Letter headline #1: The result showed that all "good" headlines were not equal. The best wording increased the response by 0.192 percentage points.

The response rate from the 400,000 control mailings was 2.1%. The average response for the test was 1.298%. The predicted response rate for the implied best strategy, starting with the overall average and adding one-half of each significant effect, amounted to 2.40%. This represented a 15% predicted increase over the response rate of the "control."

Conclusions

The divergent paths followed by experimental statisticians and marketers for decades have begun to converge. Business leaders have started to embrace efficient multivariable test methods that offer new opportunities to test more elements of the marketing mix, reduce the time and cost of testing, gain greater market intelligence, and pinpoint the most effective marketing tactics. The challenge is applying advanced techniques with proactive restraint: using manageable yet robust test designs, focusing on strategic testing to maximize return-on-investment, and balancing the quest for speed and simplicity with the need to test bold new ideas and manage the uncertainty of dynamic markets.

References

1 Bell, G.H., Ledolter J., & Swersey, A.J. (2006). Experimental design on the front lines of marketing: Testing new ideas to increase direct mail sales. *International Journal of Research in Marketing*, 23(3), 309-319.

2 Content for this section is drawn from the article: Bell, G.H., & Longbotham, R. (2007). (Sub-) optimal test designs for multivariable marketing testing. *Quirk's Marketing Research Review*, 21 (2), 20-23.

3 Van Wert, C. (with Bell, G., & Nelson, T.) speaking at the session, Multivariable Testing: Secrets of Success from Conde Nast and Ameriprise, at the *Direct Marketing Association Annual Conference (DMA06)*. October 16, 2006.

4 The content and case study that follows is taken directly from the author's earlier paper (with Ledolter, J. & Swersey, A.J.), referenced (#1) above.