Boost Your Marketing ROI with Experimental Design

by Eric Almquist and Gordon Wyner

Harvard Business Review

Reprint R0109K

Harvard Business Review

October 2001

Boost Your Marketing ROI with Experimental Design

by Eric Almquist and Gordon Wyner

Most marketing executives will admit that their discipline involves a lot of guesswork. But by borrowing a statistical technique long applied in other fields, marketers can develop campaigns that target customers with uncanny accuracy.

CONSUMERS ARE BOMBARDED daily with hundreds, perhaps thousands, of marketing messages. Delivered through all manner of media, from television commercials to telephone solicitations to supermarket circulars to Internet banner ads, these stimuli may elicit the desired response: The consumer clips a coupon, clicks on a link, or adds a product to a shopping cart. But the vast majority of marketing messages fail to hit their targets. Obviously, it would be valuable for companies to be able to anticipate which stimuli would prompt a response since even a small improvement in the browse-to-buy conversion rate can have a big impact on profitability. But it has been very difficult to isolate what drives consumer behavior, largely because there are so many possible combinations of stimuli.

Now, however, marketers have easier access, at relatively low cost, to experimental design techniques long applied in other fields such as pharmaceutical research. Experimental design, which quantifies the effects of independent stimuli on behavioral responses, can help marketing executives analyze how the various components of a marketing campaign influence consumer behavior. This approach is much more precise and cost effective than traditional market testing. And when you know how customers will respond to what you have to offer, you can target marketing programs directly to their needs–and boost the bottom line in the process.

Traditional Testing

The practice of testing various forms of a marketing or advertising stimulus isn't new. Direct marketers, in particular, have long used simple techniques such as split mailings to compare how customers react to different prices or

promotional offers. But if they try to evaluate more than just a couple of campaign alternatives, traditional testing techniques quickly grow prohibitively expensive.

Consider the "test and control cell" method, which is the basis for almost all direct mail and e-commerce testing done today. It starts with a control cell for, say, a base price, then adds test cells for higher and lower prices. To test five price points, six promotions, four banner ad colors, and three ad placements, you'd need a control cell and 360 test cells ($5 \times 6 \times 4 \times 3 = 360$). And that's a relatively simple case. In credit card marketing, the possible combinations of brands, cobrands, annual percentage rates, teaser rates, marketing messages, and mail packaging can quickly add up to hundreds of thousands of possible bundles of attributes. Clearly, you cannot test them all.

There's another problem with this brute force approach: It typically does not reveal which individual variables are causing higher (or lower) responses from customers, since most control-cell tests reflect the combined effect of more than two simple variables. Is it the lower price that prompted the higher response? The promotional deal? The new advertising message? There's no way to know.

The problem has been magnified recently as companies have gained the ability to change their marketing stimuli much more quickly. Just a few years ago, changing prices and promotions on a few cans of food in the supermarket, for example, required the time-consuming application of sticky labels and the distribution of paper coupons. Today, a store can adjust prices and promotions electronically by simply reprogramming its checkout scanners. The Internet has further heightened marketing complexity by reducing the physical constraints on pricing, packaging, and communications. In the extreme, an on-line retailer could change the prices and promotion of every product it offers every minute of the day. It can also change the color of banner ads, the tone of promotional messages, and the content of outbound e-mails with relative ease.

The increasing complexity of the *stimulus-response network,* as we call it, means that marketers have more communication alternatives than ever before – and that the portion of alternatives they actually test is growing even smaller. But this greater complexity can also mean greater flexibility in your marketing programs – if you can uncover which changes in the stimulus-response network actually drive customer behavior. One way to do this is through scientific experimentation.

A New Marketing Science

The science of experimental design lets people project the impact of many stimuli by testing just a few of them. By using mathematical formulas to select and test a subset of combinations of variables that represent the complexity of all the original variables, marketers can model hundreds or even thousands of stimuli accurately and efficiently.

This is not the same thing as an after-the-fact analysis of consumer behavior, sometimes referred to as data mining. Experimental design is distinguished by the fact that you define and control the independent variables before putting them into the marketplace, trying out different kinds of stimuli on customers rather than observing them as they have naturally occurred. Because you control the introduction of stimuli, you can establish that differences in response can be attributed to the stimulus in question, such as packaging or color of a product, and not to other factors, such as limited availability of the product. In other words, experimental design reveals whether variables *caused* a certain behavior as opposed to simply being *correlated* with the behavior.

While experimental design itself isn't new, few marketing executives have used the technique – either because they haven't understood it or because day-to-day marketing operations have gotten in the way. But new technologies are making experimental design more accessible, more affordable, and easier to administer. (For more information on the genesis of this type of testing, see the sidebar "The Origins of Experimental Design.") Companies today can collect detailed customer information much more easily than ever before and can use those data to build models that predict customer response with greater speed and accuracy.

Today's most popular experimental-design methods can be adapted and cus-

Eric Almquist *and* ***Gordon Wyner*** *are vice presidents of Mercer Management Consulting, based in Boston.*

The Origins of Experimental Design

Experimental design methodologies – some dating as far back as the nineteenth century – have been used for years across many fields, including process manufacturing, psychology, and pharmaceutical clinical trials, and they are well known to most statisticians. Sir Ronald A. Fisher was among the first statisticians to introduce the concepts of randomization and analysis of variance. In the early 1900s, he worked at the Rothamsted Agricultural Experimental Station outside London. His focus was on increasing agricultural yields.

Another major breakthrough in the field came with the work of U.S. economist and Nobel laureate Daniel L. McFadden in the 1970s, who drew on psychological theories to explain that consumer choices are a function of the available alternatives and other consumer characteristics. In helping to design San Francisco's BART commuter rail system, McFadden analyzed the way people evaluate trade-offs in travel time and cost and how those trade-offs affect their decisions about means of transportation. He was able to help forecast demand for BART and determine where best to locate stations. The model was quite accurate, predicting a 6.4% share of commuter travel for BART, which was close to the actual 6.2% share the system achieved.

tomized using guidelines from standard reference textbooks such as *Statistics for Experimenters* by George E. P. Box, J. Stuart Hunter, and William G. Hunter; and from off-the-shelf software packages such as the Statistical Analysis System, the primary product of SAS Institute. A handful of companies have already applied some form of experimental design to marketing. They include financial firms such as Chase, Household Finance, and Capital One, telecommunications provider Cable & Wireless, and Internet portal America Online.

Applying experimental-design methods requires business judgment and a degree of mathematical and statistical sophistication – both of which are well within the reach of most large corporations and many smaller organizations. The experimental design technique is particularly useful for companies that have large numbers of customers and that face rapid and constant change in their markets and product offers. Internet retailers, for instance, benefit greatly from experimentation because on-line customers tend to be fickle. Attracting browsers to a Web site and then converting them into buyers has proved very expensive and largely ineffective. Getting it right the first time is nearly impossible, so experimentation is critical. The rigorous and robust nature of experimental design, combined with the increasing challenges of marketing to oversaturated consumers, will make widespread adoption of this new marketing science only a matter of time in most industries.

The ABCs of Experimental Design

To illustrate how experimental design works, let's consider the following simple case. A company, which we'll call Biz Ware, is marketing a software product to other companies. Before launching a national campaign, Biz Ware wants to test three different variables, or *attributes,* of a sales message for the product: price, message, and promotion. Each of the three attributes can have a number of variations, or *levels.* Suppose the three attributes and their various levels are as follows:

Price

LEVEL	(1)	(2)	(3)	(4)
	$150	$160	$170	$180

Message

(1) SPEED
"Biz Ware lets you manage customer relationships in just minutes a day."

(2) POWER
"Biz Ware can be expanded to handle a virtually infinite number of customer files."

Promotion

(1) 30-DAY TRIAL
"You can try Biz Ware now for 30 days at no risk."

(2) FREE GIFT
"Buy Biz Ware now and receive our contact manager software absolutely free."

The total number of possible combinations can be determined by multiplying the number of levels of each attribute. The three attributes Biz Ware wants to test yield a total of 16 possible combinations since $4 \times 2 \times 2 = 16$. All 16 combinations can be mapped in the cells of a simple chart like the one below.

Biz Ware's Universe of Possible Combinations

PROMOTION	(1)	(1)	(2)	(2)
MESSAGE	(1)	(2)	(1)	(2)
PRICE (1)	X	X	X	X
PRICE (2)	X	X	X	X
PRICE (3)	X	X	X	X
PRICE (4)	X	X	X	X

It's not necessary to test them all. Instead, using what's called a *fractional factorial design,* Biz Ware selects a subset of eight combinations to test.

"Factorial" means Biz Ware "crosses" each attribute (price, promotion, and message) with each of the others in grid-like fashion, as in the universe chart above. "Fractional" means Biz Ware then chooses a subset of those combinations in which the attributes are independent (either totally or partially) of each other. The following chart shows the resulting experimental design, with Xs marking the cells to be tested. Note that each level of each attribute is paired in at least one instance with each level of the other attributes. So, for example, price at $150 is matched at some point with each promotion and each message. This makes it possible to unambiguously separate the influence of each variable on customer response.

Biz Ware's Experimental Design

PROMOTION	(1)	(1)	(2)	(2)
MESSAGE	(1)	(2)	(1)	(2)
PRICE (1)	X			X
PRICE (2)		X	X	
PRICE (3)		X	X	
PRICE (4)	X			X

The eight chosen combinations are now tested, using one of several media: scripts at a call center, banner ads on Biz Ware's Web site, e-mail messages to prospective customers, and direct mail solicitations. (In general, you should test using the medium you ultimately expect to use for your marketing campaign, although you can also choose multiple media and treat the choice of media as an attribute in the experiment.)

How big should the sample size be to make the experiment valid? The answer depends on several characteristics of the test and the target market. These may include the expected response rate, based on the results of past marketing efforts; the expected variation among subgroups of the market; and the complexity of the design, including the number of attributes and levels. In any event, the sample size should be large enough so that marketers can statistically detect the impact of the attributes on customer response. Since increasing the complexity and size of an experiment generally adds cost, marketers should determine the minimum sample size necessary to achieve a degree of precision that is useful for making business decisions. (There are standard guidelines in statistics that can help marketers answer the question of sample size.) We've conducted complex experiments by sending e-mail solicitations to lists of just 20,000 names, where 1,250 people each receive one of 16 stimuli.

Within a few days or weeks, the experiment's results come in. Biz Ware's marketers note the number and percentage of positive responses to each of the eight tested offers.

Biz Ware's Design Results

PROMOTION	(1)	(1)	(2)	(2)
MESSAGE	(1)	(2)	(1)	(2)
PRICE (1)	14%			40%
PRICE (2)		9%	13%	
PRICE (3)		6%	10%	
PRICE (4)	1%			7%

At a glance, you might intuitively understand that price has a significant impact on the response to the various offers, since the lower price offers (Price 1 and Price 2) generally drew much better response rates than the higher price offers (Price 3 and Price 4). But statistical modeling, using standard software, makes it possible to assess the impact of each variable with far greater precision. Indeed, by using a method known as *logistic regression analysis,* Biz Ware can extrapolate from the results of the experiment the probable response rates for all 16 cells.

Biz Ware's Modeled Responses

PROMOTION	(1)	(1)	(2)	(2)
MESSAGE	(1)	(2)	(1)	(2)
PRICE (1)	14%	23%	28%	42%
PRICE (2)	7%	12%	15%	24%
PRICE (3)	3%	6%	7%	12%
PRICE (4)	1%	3%	3%	6%

Note that the percentages shown above don't precisely match the original percentages from the test. That's because Biz Ware used the original percentages to create a general equation for estimating the results in all the cells. When the new equation is then applied to the cells already tested, the results usually vary somewhat from the original numbers. The important thing is that the tester ends up with a full set of consistent results for all possible permutations. (For more about how these calculations were made, see the sidebar "Estimating a Response Model.")

With this complete picture, it becomes clear that some combinations are far more likely to be effective than others. The combination of Price 1 ($150), Message 2 (Power), and Promotion 2 (Free Gift) is clearly the most attractive to consumers. But is it the right combination for Biz Ware? That's where business judgment comes in: The company's management will need to analyze the experiment's implications for its resources, revenue, and profitability.

Experimentation at Crayola

Let's look at an actual example of how experimental design can enhance a marketing campaign. Last year, Crayola, a division of Binney & Smith and Hallmark, launched a creative arts and activities portal on the Internet called Crayola.com. The site's target customers include parents and educators, and it sells art supplies and offers arts-and-crafts project ideas and classroom lesson plans. We conducted an experimental design to help Crayola attract people to the site and convert browsers into buyers.

Based on Crayola's experience and market knowledge, we identified a set of stimuli that could be varied to drive traffic to Crayola.com and induce purchases. One of these stimuli was an e-mail to parents and teachers. To test various components of the e-mail content and format, we relied on the best judgment of Crayola's marketing staff about the messages that were most likely to appeal to the target markets. The e-mail included five key attributes that seemed likely to affect the customer response rate, which would be measured by click-throughs to the Crayola Web site. These attributes and their related levels were:

- *Two subject lines:* "Crayola.com Survey" and "Help Us Help You."
- *Three salutations:* "Hi [user name] :-)," "Greetings!" and "[user name]."
- *Two calls to action:* "As Crayola.com grows, we'd like to get your thoughts about the arts and how you use art materials" and "Because you as an educator have a special understanding of the arts and how art materials are used, we invite you to help build Crayola.com."
- *Three promotions:* a chance to participate in a monthly drawing to win $100 worth of Crayola products; a monthly drawing for one of ten $25 Amazon.com gift certificates; and no promotion.

Estimating a Response Model

Logistic regression analysis is a statistical technique that allows an experimenter to analyze the impact of each stimulus in an experiment. The formula assumes that the outcome – in Biz Ware's case, the customer response rate – is a function of the attributes – in Biz Ware's case, price, message, and promotion. Here's what Biz Ware's generic equation looks like:

$$\text{Log}\left(\frac{\text{response rate}}{1-\text{response rate}}\right) = a + b_1\,(\text{price}) + b_2\,(\text{message}) + b_3\,(\text{promotion})$$

We plug Biz Ware's customer response data into this equation, using SAS software to estimate the coefficients (a, b_1, b_2, and b_3). For price we can drop a number into the formula. For message and promotion, which are qualitative attributes, we assign a dummy code – zero or one, since there are only two levels for each attribute. It does not matter which attribute is assigned which number. For Biz Ware, the equation looks like this:

$$\text{Log}\left(\frac{\text{response rate}}{1-\text{response rate}}\right) = 10.3 - 8.1\,(\text{price}) + 0.6\,(\text{message}) + 0.9\,(\text{promotion})$$

The coefficients tell us a few things: Higher price has a negative impact on demand (hence, the coefficient b_1 is –8.1) and the effect of promotion is greater than the effect of message (because 0.9 is greater than 0.6). But more important, these coefficients allow us to apply the equation to extrapolate from the data collected and predict responses for all 16 cells.

Crayola Draws Results

Sample E-mail

Crayola.com marketers wanted to measure how customer response would be affected by different variations (or levels) of five main e-mail attributes: two subjects, three salutations, two calls to action, three promotions, and two closings. At right is one of the 72 possible combinations. An experimental design was developed so that only 16 combinations had to be tested.

subject	Help Us Help You
salutation	Greetings!
call to action	Because you as an educator have a special understanding of the arts and how art materials are used, we invite you to help build Crayola.com. By answering ten quick questions, you'll be helping to make sure improvements we make to Crayola.com meet your needs. Simply follow this link: http://_____. (If this does not work, please cut and paste the link into your browser's address bar.)
promotion	As a thank you, you will be entered into our monthly drawing to win one of ten $25 Amazon.com gift certificates. By completing our survey, you'll be automatically entered. Thanks for your assistance. Be sure to check back often for new fun, creative, and colorful ideas and solutions.
closing	Yours, Crayola.com

Parents' Response Rates

The company measured the responses it received and through statistical modeling could quickly pinpoint which stimuli appealed most to its target customers – in this case, parents.

The subject line "Crayola.com Survey," for example, was more effective at creating positive responses than "Help Us Help You." The response rate of the former was 7.5% higher than that of the latter, all else being equal.

	Levels	Change in Response Rate
subject	Crayola.com Survey	7.5%
	Help Us Help You	0.0%
salutation	Hi [user name] :-)	2.7%
	Greetings!	0.0%
	[user name]	3.4%
call to action	As Crayola.com grows…	0.0%
	Because you are…	3.5%
promotion	$100 product drawing	8.4%
	$25 Amazon.com gift certificate drawing	5.2%
	No offer	0.0%
closing	Crayola.com	0.0%
	EducationEditor@Crayola.com	1.2%

Script Attributes

The combination of attributes that got the best response from parents was more than three times as effective as the combination of attributes that got the worst response. At right are the best and worst script attributes of the 72 possible combinations.

	Best Response	Worst Response
subject	Crayola.com Survey	Help Us Help You
salutation	User name	Greetings!
call to action	Because you are…	As Crayola.com grows…
promotion	$100 product drawing	No offer
closing	EducationEditor@Crayola.com	Crayola.com
response rate	**33.7%**	**9.7%**

The best script is 3.5 times more effective than the worst.

• *Two closings:* "Crayola.com" and "EducationEditor@Crayola.com."

Taking into account all the levels of each attribute, there were a total of 72 possible versions of the e-mail ($2 \times 3 \times 2 \times 3 \times 2 = 72$). While Crayola might have been able to test all 72 variations, the process would have been cumbersome and expensive. Instead, we created a subset of 16 e-mails to represent the 72 possible combinations. Over a two-week period, we sent the 16 types of e-mail to randomly selected samples of customers and tracked and analyzed their responses. The results were compelling. The "best" e-mail of the 72 possible scripts yielded a positive response rate of about 34% and was more than three times as effective at attracting parents as the "worst" e-mail, which yielded a positive response rate of only about 10%. (See the exhibit "Crayola Draws Results.") Among educators, the best e-mail script was nearly twice as effective as the worst script, with response rates of 35% for the best e-mail versus 20% for the worst.

We also conducted similar experiments with Crayola to test the effects of three different banner ads on the home page, as well as product, price, and promotions offered at the on-line store. The best combination was nearly four times as effective in converting shoppers into buyers as the worst combination and nearly doubled revenues per buyer.

Uncovering the Unexpected

In the Crayola experiments, as well as in other tests, our results have yielded surprising insights. When we ask experienced marketers to predict which stimuli are likely to elicit the best responses, few get it right. Crayola, for example, was surprised to find that a price reduction on a product or line of products generated sufficient volume to create higher revenues while also maintaining the site's profitability. Conventional wisdom would have suggested that raising prices would be a more effective way to increase revenue.

Of course, the findings from tests like these can't be generalized. For instance, compare the e-mail tests we conducted at Crayola with similar tests we performed for Cable & Wireless. At Crayola, e-mails containing no promotional offer drew poorly compared with e-mails containing either of the two promotional offers tested (a $100 product drawing and a $25 Amazon.com gift certificate). At Cable & Wireless, though, e-mails with no promotional offer drew the *best* click-through rate. But that's part of the value of experimental design: It allows marketers to move beyond rules of thumb or experience to pinpoint the marketing approaches that work best with a particular audience in a particular marketplace at a particular moment in time.

The Expanding Marketing Universe

In the world of marketing experimentation, the Crayola tests are relatively simple. We tested only a handful of marketing attributes, with a relatively small number of levels for each. Even so, the customer impact was impressive, with the best combinations of stimuli drawing double, triple, or quadruple response rates compared with the worst combinations.

The approach we took with Crayola can be extended and applied to more attributes and more levels. It's not unusual for a company to test ten or more attributes, including some with as many as eight levels. A credit card company, for example, might be interested in testing six teaser rates, six cobrands, four different annual percentage rates, six promotions, four insurance packages, four modes of communication, eight direct mail packages, and four mailing schedules. This represents a possible set of 442,368 distinct marketing stimuli, obviously too large a universe for a test-and-control-cell approach. But by using experimental design to select and test a manageable number – say 128 combinations of these variables – the credit card company could estimate with great accuracy the customer reaction to all 442,368 combinations.

And this is by no means the upper limit of the usefulness of experimental design. The responses being sought from customers can be more complex as well. Customers may have multiple options to choose from rather than a simple "yes" or "no" response – for example, a choice between one-year, two-year, and three-year subscriptions or no purchase.

Different types of experimental designs can be used when the experimental objectives vary. For example, so-called *screening designs* can efficiently test very large numbers of attributes to select a smaller number to investigate in more detail. Subsequent testing can employ more levels for each of a smaller collection of attributes. *Response surface designs* are used in food testing in which multiple dimensions such as sweet, salty, crunchy, and sour have an ideal level somewhere between "too much" and "not enough." The design lets testers estimate the ideal combination of tastes and textures.

Getting Results

Naturally, there are limits to the power of experimental design. This approach requires thoughtful planning to hypothesize what you are looking for and to rule out other possibilities before the experiment can begin.

One caution is that many experimental designs rely on "main effects" models. That is, they assume that interaction effects – the impact that one variable can have on another variable – are negligible. This is usually a reasonable assumption when you're dealing with complex combinations of three, four, and five variables at a time. However, interactions between two variables can be important and can be tested. For example, suppose you find that free samples have a more positive impact on a product's sales than do coupons. You may also learn that free samples provide an even greater lift when they are handed out in the stores as opposed to being sent through the mail. One variable, the chosen distribution channel, interacts meaningfully with another variable, the promotion itself.

Experimental design also calls for substantive knowledge to frame the problem, careful application of theoretically sound methods, and skillful interpretation of the results in the ap-

propriate context. Some knowledge of basic statistical methods is necessary, of course. But even more important is a nuanced understanding of customers and the ability to form reasonable assumptions concerning which attributes and levels should be tested and which shouldn't. Experimental design should be one part of a continuous test-and-learn cycle.

Marketing is, and always will be, a creative endeavor. But it doesn't have to be so mysterious. As marketing noise and advertising clutter continue to increase, marketers will find that scientific experimentation will allow them to better communicate with their customers – and substantially raise the odds that their marketing efforts will pay off.

Reprint R0109K

To place an order, call 1-800-988-0886.