ELSEVIER

Intern. J. of Research in Marketing 23 (2006) 309–319

International Journal of Research in Marketing

www.elsevier.com/locate/ijresmar

Experimental design on the front lines of marketing: Testing new ideas to increase direct mail sales

Gordon H. Bell [a], Johannes Ledolter [b,d,*], Arthur J. Swersey [c]

[a] *LucidView, 80 Rolling Links Boulevard, Oak Ridge, TN 37830, USA*
[b] *C. Maxwell Stanley Professor of Management Science, Tippie College of Business, S352 PBB, University of Iowa, Iowa City, IA 52242, USA*
[c] *Operations Research, Yale School of Management, New Haven, CT 06520, USA*
[d] *Vienna University of Economics and Business Administration, Vienna, Austria*

Abstract

Marketers have recently begun to embrace complex experimental designs for marketing and advertising testing. Full-factorial, fractional-factorial and Plackett–Burman designs have given marketers new statistical tools to increase the speed, power, and profitability of their testing programs. This case study shows how well constructed and managed experimental designs offer marketing professionals clear, bottom-line benefits over common change one variable at-a-time testing techniques.

Keywords: Direct mail; Factorial design; In-market testing; Plackett–Burman design

1. Introduction

"Test everything" has been a rallying cry in the marketing and advertising industry throughout the twentieth century. Industry experts like Claude Hopkins (1966), John Caples (1974), David Ogilvy (1983), and Bob Stone and Ron Jacobs (2001) have stressed the importance of testing new ideas in the marketplace. But as statisticians developed and refined sophisticated experimental design techniques, most marketers held firm to the change one variable at-a-time approach, often called "split-run testing" (also referred to as A/B splits, test-control, or champion–challenger testing). Only in the last few years have marketing leaders begun to embrace advanced techniques for real-world testing.

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. Tests were run within two direct mail campaigns that focused on increasing the number and profitability of new customers. The initial experiment, a Plackett–Burman screening design of 19 factors in 20 runs, was followed by a four-factor 16-run full-factorial experiment.

Although factorial, fractional factorial, and related methods of experimental design have been widely applied to manufacturing problems, there have been few applications to direct mail, Internet, retail, and other market testing programs, and we found no papers that apply Plackett–Burman designs to these problems. For in-market testing, in an early paper, Curhan (1974) used a fractional-factorial design to examine the effects of price, advertising, display space, and display location on the sales of fresh fruits and vegetables in a supermarket, while Barclay (1969) used a factorial design to evaluate the effect on profitability of raising the prices of two retail products manufactured by the Quaker Oats Company. Holland and Cravens (1973) presented the essential features of fractional-factorial designs and illustrated them with a hypothetical example concerning the effect of advertising and other factors on the sales of candy bars. Wilkinson, Wason, and Paksoy (1982) described a factorial experiment for assessing the impact of price, promotion, and display on the sales of selected items at Piggly Wiggly grocery stores.

Although the market testing literature is sparse on the use of experimental design models with many factors, one or two factor

* Corresponding author. C. Maxwell Stanley Professor of Management Science, Tippie College of Business, S352 PBB, University of Iowa, Iowa City, IA 52242, USA. Tel.: +1 319 335 3814; fax: +1 319 335 0297.
E-mail address: johannes-ledolter@uiowa.edu (J. Ledolter).

0167-8116/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijresmar.2006.05.002

experiments have been common. For example, Lodish et al. (1995a) analyzed the results of 389 television advertising experiments to determine the effect of advertising on sales. In their data set, there were three types of tests: comparing two different versions of advertising copy, comparing two different levels of exposure, and testing copy and exposure simultaneously using a factorial design. In a related paper, Lodish et al. (1995b) examined the carryover effect of television advertising exposure by tracking sales for an additional two years beyond the original one-year test period.

Factorial and fractional-factorial designs are well known and have been widely used in behavioral marketing experiments in laboratory settings (see for example, Jaffe, Jamieson, and Berger (1992), Srivastava and Lurie (2004), and Ettenson and Wagner (1986)), and in conjoint analysis applications. Green, Krieger, and Wind (2001) describe a credit card study that illustrates how fractional-factorial designs may be used in conjoint analysis. Their design consisted of 12 attributes relating to potential credit card services, each having two to six levels. For example, annual price (six alternatives), retail purchase insurance (no, yes), rental car insurance (no, yes), and airport club admission (no admission, $5 fee per visit, $2 fee per visit). Using a fractional-factorial design, 64 profiles were created out of a total of 186,624 possible attribute-level combinations. The 64 profiles were partitioned into "blocks" of eight profiles each, with all profiles in a given block being presented to each respondent. For each profile of credit card services, the respondent was asked to indicate the likelihood of purchase on a 0–100 point scale. This blocked fractional design provided independent (uncorrelated) estimates of main effects.

Green, Carroll, and Carmone (1978) provide an excellent overview and discussion of the key elements in fractional-factorial designs, while Green and Srinivasan (1978), Green and Srinivasan (1990), and Green et al. (2001) provide notable reviews of the extensive literature on conjoint analysis. Bradlow (2005) discusses current issues in conjoint analysis and the need for future research, while Wittink and Cattin (1989) and Wittink, Vriens, and Burhenne (1994) document the widespread commercial use of conjoint models. Although Green, Carroll, and Carmone (1978) briefly discuss Plackett–Burman designs, we found no papers that used these designs in conjoint and discrete choice models.

Our Plackett–Burman design is a main effect model that, as we will show, may provide evidence of likely two-factor interactions under some circumstances. The fractional designs used in conjoint analysis are typically main effects models as well, confounding main effects and two-factor interactions. Carmone and Green (1981) show how selected two-factor interactions can be included in fractional main-effects designs. Plackett–Burman and fractional-factorial models are orthogonal designs, which means that effects are estimated independently, and with minimum variance. Orthogonal designs may be prohibitively large in situations with many factors, including some at more than two levels, and in cases where interactions are important. For these circumstances non-orthogonal designs are available and may be generated using statistical software. Kuhfeld, Tobias, and Garratt (1994) discuss such non-orthogonal designs and their use in conjoint and discrete choice studies.

Our review of the literature shows that fractional designs and related orthogonal designs have been used extensively in conjoint and discrete choice studies. As we have noted, there have also been a few papers on market tests involving relatively few factors that use factorial or fractional-factorial designs. However, it has been our experience that until recently the great majority of market testing practitioners relied on the traditional approach of testing one factor at-a-time. In this paper we show the benefits of statistical methods that simultaneously test many factors, and demonstrate the usefulness of Plackett–Burman designs, an important class of experimental design models.

2. The experiment

2.1. The factors

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 Table 1 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 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. The firm's marketing staff believed that a sticker increases involvement and is likely to increase the number of orders. 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.

2.2. A Plackett–Burman design for 19 factors

With so many factors, we chose a two-level design. By doing so, we could keep the number of runs relatively low and avoid

Table 1
The 19 test factors and their low and high levels

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

more complicated and possibly non-orthogonal designs. Two-level screening designs are common in the field of experimental design; see Box, Hunter, and Hunter (1978). Our philosophy in testing many factors, each at two levels, was to identify which factors were active, i.e., had a significant effect on the response. Once these active factors were identified, it would be possible, if needed, to test each of them at more than two levels, while still maintaining an orthogonal design.

With 19 factors, we created the 20-run *Plackett–Burman main effects design* shown in Table 2. Plackett and Burman (1946) designs are orthogonal designs for factors that have two levels each, with the number of runs N given by a multiple of 4. For two-level fractional factorials the run size N must be a power of 2, leaving large gaps in the run sizes. For example, a minimum of 32 runs is required in a fractional-factorial design involving 19 factors. The Plackett–Burman design, on the other hand, can study 19 factors in just 20 runs. This is why these designs are useful in situations where the number of runs is critical.

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 runs. In the 20-run design (Table 2), 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 Table 2, 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.

In what follows, we will assume that three-factor and higher order interactions are negligible and therefore can be ignored. The main effect of a factor is the difference between the response averages at the high (plus) and low (minus) levels of that factor. Both fractional-factorial designs and Plackett–Burman designs are orthogonal, but the natures of their confounding patterns differ. Consider a fractional-factorial design in which main effects are confounded with two-factor interactions; for example, the saturated design for 15 factors in 16 runs shown in Table 3. The design matrix is constructed by first writing columns of signs for a full-factorial design in four factors (columns A–D). The signs for the remaining columns are determined from eleven generators that use all interaction columns in the full-factorial design. For example, consider the generator K = ABC. Multiplying the signs in columns A, B, and C, row by row, results in the column of signs for factor K. There are 15 main effects and 105 two-factor interactions (15!/2!13!). Each interaction belongs to a single set of seven two-factor interactions, and each main effect is confounded with one of these sets. For example, we find that A is confounded with BE, CF, DG, HK, IL, JM, and NO. The factor A does not appear as a letter in any of the seven interactions, and no two interactions include the same factor. The column of signs for factor A is identical to the column of signs for each of the two-factor interactions that are confounded with the main effect of A. Hence there is perfect correlation ($\rho = 1$) between the column of signs of A and the column of signs for each of its confounded two-factor interactions. For example, multiplying the signs in columns B and E row by row to obtain a column representing the BE interaction results in a column of signs that is identical to the column of signs for factor A. Because of this perfect correlation, estimating the main effect by taking the difference between the response averages at the high (plus) and low (minus) levels of a particular factor, actually gives an estimate of the main effect of that factor *plus* the sum of the seven two-factor interactions that are confounded with that main effect. If all of these interactions are negligible, the result will be a clear estimate of the main effect. If one or more of the interactions are significantly different from zero, the estimate of the main effect will be biased. The books by Berger and Maurer (2002), Box et al. (1978), and Ledolter and Burrill (1999) discuss fractional-factorial designs, confounding, and the analysis of experimental results.

Plackett–Burman designs have more complex confounding patterns. Each main effect is confounded with all two-factor interactions except those that involve that main effect. In our 19 factor design in Table 2, the main effect for each factor is confounded with all two-factor interactions involving the other 18 factors, a total of 153 interactions (18!/2!16!). But in contrast to the fractional-factorial design shown in Table 3, the column of signs for each main effect is not identical to the column of signs for each of its confounded two-factor interactions. Although not identical, and therefore not perfectly correlated, these columns of signs are correlated. That is, the correlation between the signs in a main effect column and the signs in each two-factor interaction column that is confounded with that main effect, is strictly less than 1 in absolute value ($|\rho| < 1$). As a consequence, it can be shown (see Chapter 6 of Ledolter and Swersey (in press)) that estimating the main effect of a particular factor by taking the difference between the high (plus) and low (minus) levels for that factor actually provides an estimate of the main effect plus the *weighted* sum of the two-factor interactions that are confounded with that main effect. The weight associated with each two-factor interaction is the correlation between that two-factor interaction and the main effect; see Barrentine (1996) for a discussion of the structure of confounding patterns in Plackett–Burman designs. By enumerating all correlations among factor columns and interaction columns we find that for the 20-run Plackett–Burman design in Table 2, the weights (correlations) are either −0.2, +0.2 or −0.6. Of the 153 interactions confounded with each main effect, 144 have weights of −0.2 or +0.2, while nine interactions have weights of −0.6. A particular two-factor interaction will appear in the confounding pattern of 17 main effects. For 16 of these main effects, the weight associated with this interaction will be −0.2 or +0.2, while for a single main effect, the weight associated with this interaction will be −0.6. For example, consider the main effect of factor R and the SG interaction. We use +1 and −1 to represent the column signs and multiply the entries in

Table 2
Response rates in the 20-run Plackett–Burman design

	Envelope teaser	Return address	"Official" ink-stamp on envelope	Postage	Additional graphic on envelope	Price graphic on letter	Sticker	Personalize letter copy	Copy message	Letter headline
Test cell	A	B	C	D	E	F	G	H	I	J
1	+	+	−	−	+	+	+	+	−	+
2	−	+	+	−	−	+	+	+	+	−
3	+	−	+	+	−	−	+	+	+	+
4	+	+	−	+	+	−	−	+	+	+
5	−	+	+	−	+	+	−	−	+	+
6	−	−	+	+	−	+	+	−	−	+
7	−	−	−	+	+	−	+	+	−	−
8	−	−	−	−	+	+	−	+	+	−
9	+	−	−	−	−	+	+	−	+	+
10	−	+	−	−	−	−	+	+	−	+
11	+	−	+	−	−	−	−	+	+	−
12	−	+	−	+	−	−	−	−	+	+
13	+	−	+	−	+	−	−	−	−	+
14	+	+	−	+	−	+	−	−	−	−
15	+	+	+	−	+	−	+	−	−	−
16	+	+	+	+	−	+	−	+	−	−
17	−	+	+	+	+	−	+	−	+	−
18	−	−	+	+	+	+	−	+	−	+
19	+	−	−	+	+	+	+	−	+	−
20	−	−	−	−	−	−	−	−	−	−

columns S and G to obtain the entries in column SG. Writing each column as a row to save space, and listing the run numbers above the entries, we have

Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Column R	+1	+1	−1	−1	−1	−1	+1	−1	+1	−1	+1	+1	+1	+1	−1	−1	+1	+1	−1	−1
Column SG	−1	+1	+1	+1	+1	−1	−1	−1	−1	+1	+1	−1	−1	−1	+1	+1	−1	−1	+1	+1

Both columns have 10 plus and 10 minus signs and the entries in each column add to zero. Furthermore, the sum of the squares of the entries in each column is 20, the number of runs N. The columns are correlated. In four of the 20 runs the signs match, while in 16 runs the signs are opposite. The correlation between these two mean zero columns (call them x and z) is given by

$$\rho = \frac{\sum x_i z_i}{\sqrt{\sum x_i^2}\sqrt{\sum z_i^2}} = \frac{-12}{20} = -0.6.$$

For simplicity, suppose a single two-factor interaction confounded with a particular main effect is important. A total of 17 main effects will be confounded with that interaction. For each of these main effects, taking the difference between the high (plus) and low (minus) levels for that factor provides an estimate of the main effect plus α times the magnitude of the confounded two-factor interaction. As noted above, for 16 main effects the fraction α will be -0.2 or 0.2, and the bias in our estimate of each main effect will be relatively small — plus or minus 0.2 times the magnitude of the interaction. For a single main effect, α will be -0.6, and the bias will be -0.6 times the magnitude of the interaction.

Given the complex confounding patterns of Plackett–Burman designs, it may seem at first glance that they would not provide any useful information about 2-factor interactions. In fact, traditionally they have been used as main effects designs. But more recently Plackett–Burman designs have received much greater attention from researchers because of what Box et al. (1978, second edition 2005) call "their remarkable projective properties." In analyzing the results of our experiment in the remainder of this paper, we will discuss these projective properties and show how they can be used in certain circumstances to estimate one or more 2-factor interactions from the results of a Plackett–Burman experiment.

2.3. The results

The focus of the experiment was on increasing response rate: the fraction of people who respond to the offer. A large mailing list of potential customers was available for the test. 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 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 5000 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

List of benefits	Postscript on letter	Signature	Product selection	Value of free gift	Reply envelope	Information on buckslip	2nd buckslip	Interest rate	Orders	Response rate
K	L	M	N	O	P	Q	R	S		
−	+	−	−	−	−	+	+	−	52	1.04%
+	−	+	−	−	−	−	+	+	38	0.76%
−	+	−	+	−	−	−	−	+	42	0.84%
+	−	+	−	+	−	−	−	−	134	2.68%
+	+	−	+	−	+	−	−	−	104	2.08%
+	+	+	−	+	−	+	−	−	60	1.20%
+	+	+	+	−	+	−	+	−	61	1.22%
−	+	+	+	+	−	+	−	+	68	1.36%
−	−	+	+	+	+	−	+	−	57	1.14%
+	−	−	+	+	+	+	−	+	30	0.60%
+	+	−	−	+	+	+	+	−	108	2.16%
−	+	+	−	−	+	+	+	+	39	0.78%
+	−	+	+	−	−	+	+	+	40	0.80%
+	+	−	+	+	−	−	+	+	49	0.98%
−	+	+	−	+	+	−	−	+	37	0.74%
−	−	+	+	−	+	+	−	−	99	1.98%
−	−	−	+	+	−	+	+	−	86	1.72%
−	−	−	−	+	+	−	+	+	43	0.86%
+	−	−	−	−	+	+	−	+	47	0.94%
−	−	−	−	−	−	−	−	−	104	2.08%

orthogonal, the same 100,000 people are used to obtain independent estimates of each main effect.

The marketing team regularly used 25,000–50,000 names for each split-run test, so the sample size of 100,000 for the designed experiment was not much different from what had been done in the past. Power calculations using the Minitab software convinced the authors that this sample size was large enough to detect meaningful differences. Determining the statistical significance of each main effect is equivalent to the standard statistical test for comparing two independent sample proportions, in this case of size 50,000 each. The firm estimated an average response rate of 1% and wanted to be quite confident of detecting a change of 0.20% (either an increase from 1% to 1.2% or a decrease from 1% to 0.8%). At the 5% significance level with a sample of 100,000, the detection probabilities (statistical power) were found to be 0.86 for a change to 1.2% and 0.92 for a change to 0.8%. Thus with a sample size of 100,000, the authors and marketing team were very confident of being able to detect very small, but economically meaningful differences.

Table 3
A fractional-factorial design with generators E=AB, F=AC, G=AD, H=BC, I=BD, J=CD, K=ABC, L=ABD, M=ACD, N=BCD, O=ABCD

	Factors														
Run	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
2	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
3	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
4	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
5	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−
6	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
8	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
9	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
11	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
12	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
13	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
14	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
15	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Testing all factors simultaneously has very large sample size advantages compared to testing each of the 19 factors, one at-a-time. Suppose we kept the total sample size at 100,000. Then a sample of 5263 persons would be used for each of the 19 one factor at-a-time tests. Since the control package was already being sent to 400,000 people, the group of 5263 people would receive a mailing of the control with one factor changed. The two sample proportions would then be compared to determine the effect of changing that one factor. We calculated the power of such a test, using the same assumptions as above—a 5% significance level, and an average response rate of 1%. The statistical power (detection probability) is 0.32 for a change to 1.2%, and 0.29 for a change to 0.8%, compared to 0.86 and 0.92 respectively for the Plackett–Burman design as calculated above. We found that to obtain the same statistical power as the Plackett–Burman design would require a sample size of about 25,000 for each of the 19 one factor at-a-time tests, for a total sample size of 475,000 people, an increase of 375,000 persons.

2.4. Initial analysis of the results

The estimated effects, which are differences between average responses at the plus and minus levels of the factor columns, are shown in Fig. 1. 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

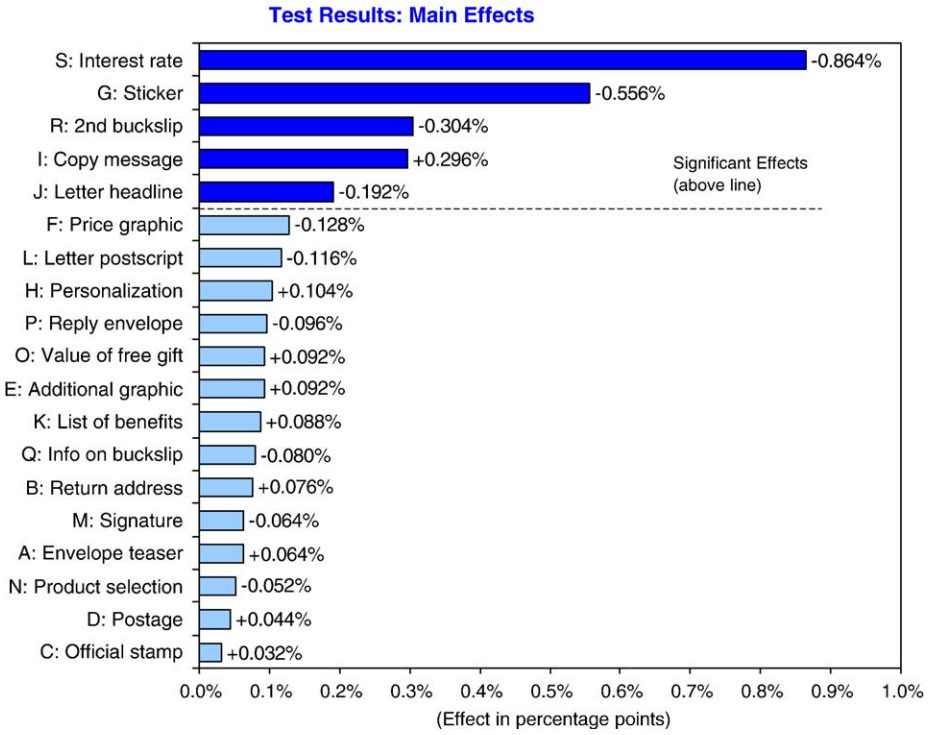

Fig. 1. Main effects estimates: Plackett–Burman design.

positive effects, the "+" level increases response; for negative effects, the "−" level increases response.

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 $n=5000$ 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

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

Replacing the unknown proportion π by the overall success proportion (averaged over all runs and samples), $\bar{p}=(\#\text{Purchases})/(nN)=1298/100{,}000=0.01298$, leads to the standard error of an estimated effect,

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

The standard error is 0.072 if effects are expressed in percentage terms. Significance (at the 5% 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 Fig. 1 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. In addition, it was very clear based on the firm's financial models that the 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 buckslip: A main effect interpretation shows that adding another buckslip reduces the number of buyers by 0.304 percentage points. One explanation offered for this surprising result was that the buckslip added unnecessary information and obscured the simple "buy now" offer. A more compelling explanation that we discuss in the next section is that the significant effect is not the result of the main effect of factor R, but is due to an interaction between two other factors.
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 the 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."

2.5. Further analysis of the results

The confounding of main effects and interactions introduces some uncertainty into our interpretation of the results. A straightforward approach for obtaining unconfounded main effects is a "foldover" of the original Plackett–Burman design. In such a foldover design the 20-run Plackett–Burman design would be augmented by an additional 20 runs, in which the signs of each of the 19 design columns are switched. The combination of a Plackett–Burman design and its complete foldover creates a design in which main effects are no longer confounded with two-factor interactions. In our experiment, a foldover was not carried out (with 40 runs it would have greatly increased the operational complexity of the mailing), and we cannot be certain which combinations of main effects and interactions are responsible for the significant estimates in Fig. 1.

The use of our Plackett–Burman design is supported by empirical experimental design principles. Effect sparsity (Box & Meyer, 1986) means that the number of important effects is typically small, while hierarchical ordering means that important interactions are usually fewer in number and smaller in magnitude compared to main effects (Wu & Hamada, 2000). In addition, on the basis of effect heredity (Hamada & Wu, 1992), the principle that significant interactions are likely to involve factors with significant main effects, it is possible in some circumstances to identify likely two-factor interactions.

Table 4a
Regression results for models relating the response rate to factors S (interest rate), G (sticker), R (2nd buckslip), I (copy message), and J (letter headline)

(a) Regression of response rate on S, G, R, and their interactions

Predictor	Coefficients	StdError	*t*-ratio	*P*-value
Constant	1.325	0.066	20.07	0.000
S	−0.386	0.066	−5.85	0.000
G	−0.320	0.066	−4.85	0.000
R	−0.061	0.066	−0.93	0.372
SG	0.151	0.066	2.29	0.041
SR	−0.070	0.066	−1.06	0.310
GR	0.076	0.066	1.16	0.271
SGR	0.045	0.066	0.68	0.508

Rate = 1.325 − (0.386)S − (0.320)G − (0.061)R + 0.151(SG) − (0.070)SR + (0.076)GR + (0.045)SGR; R^2 = 0.902.

Table 4(b)

(b) Regression of response rate on S, G, and SG

Predictor	Coefficients	StdError	*t*-ratio	*P*-value
Constant	1.298	0.052	24.75	0.000
S	−0.432	0.052	−8.24	0.000
G	−0.278	0.052	−5.30	0.000
SG	0.188	0.052	3.58	0.002

Rate = 1.298 − (0.432)S − (0.278)G + (0.188)SG; R^2 = 0.872.

Table 4(c)
Regression of response rate on S, G, SG, I, and J

Predictor	Coefficients	StdError	*t*-ratio	*P*-value
Constant	1.298	0.044	29.46	0.000
S	−0.432	0.044	−9.80	0.000
G	−0.278	0.044	−6.31	0.000
SG	0.151	0.046	3.29	0.005
I	0.118	0.045	2.62	0.020
J	−0.066	0.045	−1.46	0.166

Rate = 1.298 − (0.432)S − (0.278)G + (0.151)SG + (0.118)I − (0.066)J; R^2 = 0.921.

Factors S (interest rate) and G (presence of a sticker) are by far the largest effects in Fig. 1. The correlation between the main effect of R (2nd buckslip) and the SG interaction is −0.6. Hence a significant SG interaction would bias the estimate of the main effect of R by −0.6 times the value of the interaction. This suggests that it may not be the main effect of factor R that is important, but the two-factor interaction between S and G. This interpretation is supported by the principle of effect heredity as the main effects of S and G are the most important factors. As one might expect, at the high interest rate the effect of having a sticker is small (a change from 0.776% to 0.956% is implied by the results in Table 2), but at the low interest rate, the effect of having the sticker is much larger (a change from 1.264% to 2.024%). The sticker is most effective when the customer receives a more attractive offer.

Box and Tyssedal (1996) showed that the 20-run Plackett–Burman design produces for any three factors a complete factorial arrangement, with some combinations replicated. The design is said to have "projectivity" three. In contrast, fractional-factorial designs that confound main effects with two-factor interactions such as the one shown in Table 3 fail to produce a complete factorial for some sets of three factors, and hence only have projectivity two.[1] We use this projectivity idea to provide more evidence that the apparent main effect of R (2nd buckslip) is actually a consequence of the bias created by the SG interaction. Consider the three factors S, G, and R. Of the 20 runs in Table 2 there is at least one run at each of the eight factor-level combinations of these three factors. In specifying each combination, we let the first sign indicate the level of S, the second sign represent the level of G, and the last sign represent the level of R. There are four runs at each of the four combinations (− − −), (− + +), (+ + −), (+ − +), and one run at each of the remaining four combinations. Because we have at least one response at each combination, we have a full-factorial arrangement in factors S, G, and R (ignoring the other factors). Because the number of runs at each combination is not the same, we must use regression to estimate the effects. Doing so,

[1] A design with k factors each at two levels is said to be of projectivity p if every subset of p factors out of the possible k contains a complete 2^p factorial design, possibly with some points replicated.

Table 5
Factors and their low and high levels in the follow-up experiment

Factor	(−) Control	(+) New idea
A: Annual fee	Current	Lower
B: Account-opening fee	No	Yes
C: Initial interest rate	Current	Lower
D: Long-term interest rate	Low	High

we find that the three significant effects are S, G, and SG, confirming that it is the SG interaction and not the main effect of R that is significant.

Table 4(a) shows the results when regressing the response rate on the main and interaction effects of the three factors S, G, and R. The standard errors of the estimated regression coefficients use the pooled variance from the eight factor-level combinations, assuming that the other factors have no effect on the response. The *t*-ratios and the probability values of the regression coefficients listed in this table indicate that S, G and SG are significant, while all other effects (including the main effect of factor R) are insignificant. Table 4(b) lists the results of the regression on the significant effects S, G and SG. The regression explains 87.2% of the variability in the response rate.

Cheng (1995) showed that in the 20-run Plackett–Burman design, for any four factors, estimates of the four main effects and the six 2-factor interactions involving these four factors can be obtained when their higher-order (3- and 4-factor) interactions are assumed to be negligible. Having eliminated factor R, we apply Cheng's finding and consider a model that includes the four factors that were significant in our initial main effects analysis: S, G, I, and J, together with their six 2-factor interactions. The result of this regression shows that all 2-factor interactions except SG are insignificant, leading to a model with the four main effects and the SG interaction. The fitting results for the model with S, G, SG and the two main effects of I and J are shown in Table 4(c). The five effects explain 92.1% of the variation, a rather modest improvement over the 87.2% that is explained by S, G and SG. It is clear that factors S (interest rate) and G (sticker) and their interaction SG are the main drivers of the response rate.

3. A follow-up experiment

3.1. Full-factorial design in four factors

In light of the positive Plackett–Burman test results, the chief marketing executive wanted to continue testing. Since the long-term interest rate was such an important factor in the first test, he decided to focus on a smaller test of just interest rates and fees. In the first test, the introductory interest rate was fixed. Now, he wanted to test changes in both introductory and long-term rates, as well as the effects of adding an account-opening fee and lowering the annual fee. The four factors are shown in Table 5. Although the account-opening fee was likely to reduce response, one manager thought the fee would give an impression of exclusivity that would mitigate the magnitude of the response decline. The team also wanted once again to test the effect of a small increase in the long-term interest rate. At the same time, they wanted to test the effect of two alternative initial interest rates, both lower than the long-term rate.

Because each of the factors impacted the cost to the customer, it was expected that two-factor interactions might well exist. In order to study these interactions along with all main effects, the authors recommended a full-factorial design. The marketing team used columns A–D of the test matrix in Table 6 to create the 16 mail packages. The +/− combinations in the 11 interaction (product) columns are used solely for the statistical analysis of the results. All pairs of columns in Table 6 are orthogonal. All 15 effects (4 main effects and 11 interactions) can be analyzed independently, and none of these effects are confounded.

Table 6
Results of the follow-up experiment

	Annual fee	Account-opening fee	Initial interest rate	Long-term interest rate	(Interactions)												
Test cell	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD	Orders	Response rate
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+	184	2.45%
2	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−	252	3.36%
3	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−	162	2.16%
4	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+	172	2.29%
5	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−	187	2.49%
6	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+	254	3.39%
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+	174	2.32%
8	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−	183	2.44%
9	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−	138	1.84%
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+	168	2.24%
11	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+	127	1.69%
12	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−	140	1.87%
13	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+	172	2.29%
14	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−	219	2.92%
15	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−	153	2.04%
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	152	2.03%

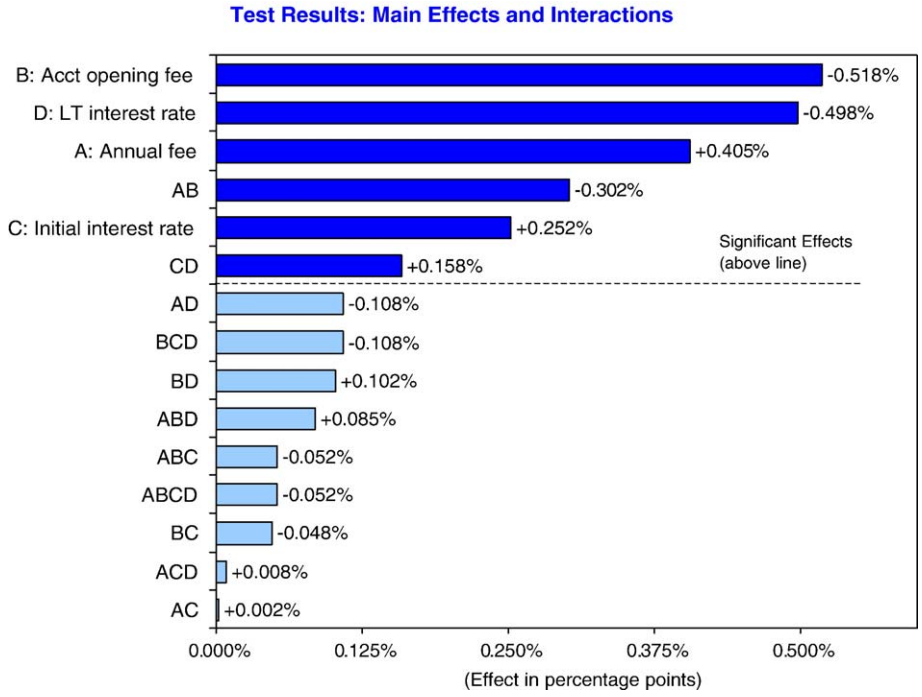

Fig. 2. Main and interaction effects: Follow-up experiment.

3.2. The results

Each of the $N=16$ test cells was mailed to $n=7500$ potential customers. A total of 2837 customers, or $100(2837)/(16)(7500)=2.364\%$, responded to the offer and placed an order. Main and interaction effects were calculated by applying the plus and minus signs to the response column and dividing the weighted sum by $N/2=8$. The results are shown in the diagram in Fig. 2. Standard errors of the effects (expressed in percent changes) are obtained by substituting $\bar{p}=0.02364$ into

$$\text{StdError(Effect)} = 100\sqrt{4/N}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.0877.$$

Effects outside $\pm 1.96(0.0877)=\pm 0.172$ are statistically significant at the 5% level.

As shown in Fig. 2, all four main effects and one, and perhaps two interactions (the AB and the CD interactions) are significant. Note that the CD interaction is just slightly smaller than 1.96 times the standard error.

B- or No account-opening fee

Although one manager had thought that charging an initial fee would give the impression of exclusivity, this fee had the largest negative effect, reducing response rate by 0.518 percentage points.

D- or Low long-term interest rate

Another attempt to slightly increase the interest rate showed, once again, that the long-term interest rate had to stay low. Raising the interest rate reduced response on average by 0.498 percentage points.

A+ or Lower annual fee

The annual fee was not charged until the end of the first year, but the fee was stated in the mailing. Not surprisingly, as with the other charges, a lower fee was better, increasing response by 0.405 percentage points.

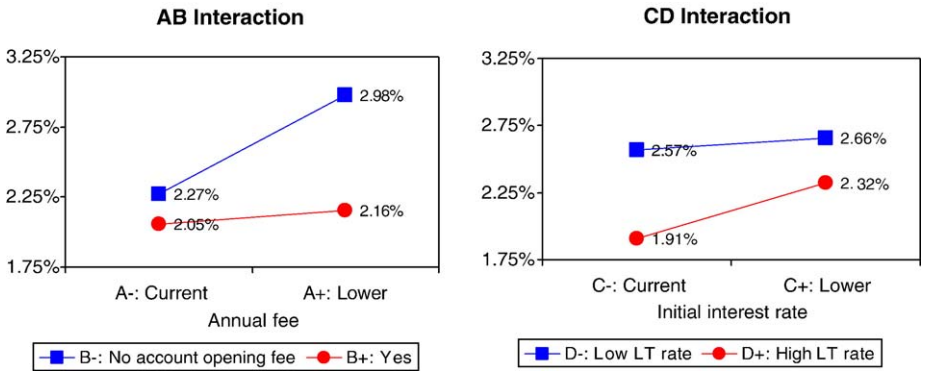

Fig. 3. Interaction plots: Follow-up experiment.

C+ or Lower initial interest rate

Reducing the introductory interest rate increased response by 0.252 percentage points.

The main effects are quite strong. However, the significant interactions (AB and CD) imply that one needs to look at the effects of A and B, and C and D jointly. The diagrams in Fig. 3 show the nature of the interactions. The AB interaction supports both main effects, but provides additional important insights. With an account-opening fee (B+), the lower annual fee results in only a small increase in response from 2.05% to 2.16%, but with no account-opening fee (B−), a lower annual fee results in a very large increase in response from 2.27% to 2.98%. The estimated response of 2.98% is highest for the combination A+B−, with both the lower annual fee and no account-opening fee. The AB interaction expresses the fact that A+ and B− together increase the response rate beyond what can be expected by each of the two factors separately. This may result from positive synergies, or may be due to the negative impact of the account-opening fee, which for some customers may cause an immediate rejection of the offer. The nature of this two-factor interaction provides extremely valuable information. Using its financial models, the company found that the increase in response resulting from no account-opening fee and a lower annual fee (A+B−) was much greater than the loss in revenue that would result from eliminating these fees.

The CD interaction shows that when the long-term rate is low (D−), the effect of a lower initial rate is small and not statistically significant (a change in response from 2.57% to 2.66%). It is clear that offering the lower initial rate would not be profitable if the lower long-term rate were also offered. However, when the long-term rate is high (D+) the lower initial rate has a large impact, with the response changing from 1.91% to 2.32%. The interaction shows that for persons receiving both lower rates, the increase in response is considerably less than the sum of the two main effects. This customer behavior is consistent with the concave value function used by Thaler (1985), and based on the earlier work of Kahneman and Tversky (1979). In contrast to the main effects that suggest both interest rates should be low, these results followed by additional analysis using the company's financial models showed that a lower long-term rate coupled with the current (higher) initial rate was the most profitable.

4. Final comments

After these two mailings – one with a 19-factor Plackett–Burman screening test and the second with the four-factor full-factorial follow-up test – the marketing team learned more than they had ever before when using simple one variable at-a-time techniques. The specific findings of these experiments led to immediate and substantial improvements: increased response rates, lower costs, and higher profits. But the longer-term benefits have been even more substantial. This study introduced the company to the use of formal experimental design methods. Since then, the firm has continued to experiment, increasing the speed and profitability of its testing programs, and becoming a leader in the application of these tools to direct marketing. Testing has given the company the ability to quickly prove what sells, and greatly improve performance in the highly competitive financial services marketplace.

Although the focus of this paper has been on direct marketing, the potential applications of experimental design approaches are widespread. Website design, online advertising, telemarketing, catalog design, and retail tests are fertile areas for multivariable experiments. As marketing applications of large fractional-factorial and Placket–Burman designs are more widely disseminated, the real-world use of these powerful techniques should become more commonplace.

Acknowledgments

We would like to thank the editor, three anonymous reviewers, and Professors Subrata Sen and K. Sudhir of the Yale School of Management for helpful comments which improved our paper considerably.

References

Barclay, W. D. (1969). Factorial design in a pricing experiment. *Journal of Marketing Research*, *6*(4), 427−429.

Barrentine, L. B. (1996). Illustration of confounding in Plackett–Burman designs. *Quality Engineering*, *9*(1), 11−20.

Berger, P. D., & Maurer, R. E. (2002). *Experimental design with applications in management, engineering and the sciences.* Belmont, CA: Duxbury Press.

Box, G.E.P., Hunter, W.G., & Hunter, J.S. (1978). *Statistics for experimenters* (2nd ed., 2005). New York: Wiley.

Box, G. E. P., & Meyer, R. D. (1986). Dispersion effects from fractional designs. *Technometrics*, *28*(1), 19−27.

Box, G. E. P., & Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika*, *83*(4), 950−955.

Bradlow, E. T. (2005). Current issues and a wish list for conjoint analysis. *Applied Stochastic Models in Business and Industry*, *21*(4/5), 319−323.

Caples, J. (1974). *Tested advertising methods* (4th ed.) Englewood Cliffs, NJ: Prentice-Hall.

Carmone, F. J., & Green, P. E. (1981). Model misspecification in multiattribute parameter estimation. *Journal of Marketing Research*, *18*(1), 87−93.

Cheng, C. S. (1995). Some projection properties of orthogonal arrays. *Annals of Statistics*, *23*(4), 1223−1233.

Curhan, R. C. (1974). The effects of merchandising and temporary promotional activities on the sales of fresh fruits and vegetables in supermarkets. *Journal of Marketing Research*, *11*(3), 286−294.

Ettenson, R., & Wagner, J. (1986). Retail buyers' saleability judgments: A comparison of information use across three levels of experience. *Journal of Retailing*, *62*(1), 41−63.

Green, P. E., Carroll, J. D., & Carmone, F. J. (1978). Some new types of fractional factorial designs for marketing experiments. In J. N. Sheth (Ed.), *Research in marketing*, *Vol. 1* (pp. 99−122).

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, *31*(Supplement 3), 56−73.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*(2), 103−123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3−19.

Hamada, M., & Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, *24*(3), 130−137.

Holland, C. W., & Cravens, D. W. (1973). Fractional factorial designs in marketing research. *Journal of Marketing Research*, *10*(3), 270−276.

Hopkins, C. C. (1966). *Scientific advertising.* Chicago: NTC Business Books original work published 1923, New York: Lord and Thomas.

Jaffe, L. J., Jamieson, L. F., & Berger, P. D. (1992). Impact of comprehension, positioning, and segmentation on advertising response. *Journal of Advertising Research*, *32*(3), 24−33.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263−292.

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, *31*(4), 545−557.

Ledolter, J., & Burrill, C. W. (1999). *Statistical quality control: Strategies and tools for continual improvement.* New York: Wiley.

Ledolter, J., and Swersey, A.J. (in press). *Testing 1–2– 3: Experimental Design for Marketing and Service Operations*. Stanford: Stanford University Press.

Lodish, L. M., Abraham, M. M., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M. E. (1995). A summary of fifty-five in-market experimental estimates of the long-term effect of TV advertising. *Marketing Science*, *14* (3), G133−G140.

Lodish, L. M., Abraham, M. M., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M. E. (1995). How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments. *Journal of Marketing Research*, *32*(2), 125−139.

Ogilvy, D. (1983). *Ogilvy on advertising.* New York: Random House.

Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, *33*(4), 305−325.

Srivastava, J., & Lurie, N. (2004). Price-matching guarantees as signals of low store prices: Survey and experimental evidence. *Journal of Retailing*, *80*(2), 117−128.

Stone, B., & Jacobs, R. (2001). *Successful direct marketing methods* (7th ed.) New York: McGraw-Hill.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*(3), 199−214.

Wilkinson, J. B., Wason, J. B., & Paksoy, C. H. (1982). Assessing the impact of short-term supermarket strategy variables. *Journal of Marketing Research*, *19*(1), 72−86.

Wittink, D. R., & Cattin, P. (1989). Commercial use of conjoint analysis: An update. *Journal of Marketing*, *53*(3), 91−96.

Wittink, D. R., Vriens, M., & Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, *11*(1), 41−52.

Wu, C. F. J., & Hamada, M. (2000). *Experiments: Planning, analysis, and parameter design optimization.* New York: Wiley.