

A Plackett-Burman Experiment to Increase Supermarket Sales of a National Magazine

Gordon H. Bell

LucidView, Oak Ridge, Tennessee 37830, gbell@lucidview.com

Johannes Ledolter

Tippie College of Business, University of Iowa, Iowa City, Iowa 52242, and
Vienna University of Economics and Business Administration, 1090 Vienna, Austria,
johannes-ledolter@uiowa.edu

Arthur J. Swersey

Yale School of Management, New Haven, Connecticut 06520, arthur.swersey@yale.edu

This paper describes and discusses a Plackett-Burman experiment aimed at increasing supermarket sales of a top-selling national magazine. The experiment involved 10 factors relating to in-store advertising and the location of the magazine within the store. We discuss issues including choice of factors, alternative designs, sample-size determination (number of test stores and the length of the test period), and the analysis of the resulting data. We show the large benefits that accrued from our approach of testing many factors simultaneously compared to the common industry practice of sequentially testing one factor at a time. We also discuss the potential opportunities of applying this approach to other service problems.

Key words: multifactor experiments; Plackett-Burman designs; retail testing; supermarket operations.

History: Published online in *Articles in Advance* March 4, 2009.

Experimental design methods have long been recognized as an integral part of production and operations management in general, and quality management in particular. With its origins in the pioneering work of Sir Ronald Fisher, who published *The Design of Experiments* originally in 1935 (Fisher 1966), experimental design methods have been widely applied to manufacturing problems, with numerous case studies and examples appearing in the literature.

In the early 1980s, largely in response to competition from Japan, US firms took a renewed interest in these statistical methods, with the Big Three automobile makers at the forefront of these activities. Experimental design was emphasized throughout that decade as an important aspect of statistical process control (SPC) and total quality management (TQM) activities. More recently, Six Sigma programs have gained widespread attention with experimental design as a prominent part of the methodology.

With respect to service operations, experimental design applications have consisted mainly of

marketing research in retail environments. These studies have focused on measuring the sales impact of price, advertising, and product displays. The opportunities to experiment in supermarket settings are particularly attractive because of small profit margins and intense competition among supermarkets. Small improvements in these operations can have major impacts on financial performance.

In this paper we discuss the design and results of an experiment aimed at increasing supermarket sales of a popular national magazine. The experiment took place in 48 stores over a two-week period and employed a Plackett-Burman design to test 10 factors that related to in-store advertising and promotion, as well as the location and display of the magazine. Each factor was tested at two possible levels (e.g., discount on the purchase of two copies or no discount, additional magazine display in the produce area or no additional display at produce, and so forth). The design consisted of 24 experimental runs; each run specified a particular combination of

settings for the 10 factors. We assigned a different pair of stores to each run, and the settings of the factors in the run determined the experimental setup for those stores during the two-week test period. The response variable (performance measure) was the percentage change in weekly sales for each pair of stores. The adopted Plackett-Burman design allowed us to estimate all main effects clear of any confounded two-factor interactions and provided evidence that two-factor interaction effects were negligible.

In comparison to the Plackett-Burman design that was implemented, a comparable two-level fractional factorial design would have consisted of 32 runs. With two stores per run, we would have required 64 stores rather than the 48 that we included in our design. Because this would have greatly increased the cost of implementing and monitoring the experiment, it would not have been feasible, given that the supermarket chain constrained the magazine publisher to limit the experiment to no more than 50 supermarket stores.

Our paper contributes to the literature in a number of ways. In our experience, in-store testing is rather common. Instead of testing multiple factors simultaneously and employing statistical models, such as completely randomized or randomized block designs, factorial designs, fractional factorial designs, Plackett-Burman designs and so forth, practitioners typically perform sequential tests of one factor at a time. These tests are usually referred to as champion/challenger or A/B tests. However, as we show in this paper, compared to designs that simultaneously test multiple factors, these one-factor-at-a-time tests require much larger sample sizes—in our case, the inclusion of more individual stores or a longer test period—and provide less information. In addition, in contrast to our 10-factor test, the in-store experiments that have been described in the academic literature generally involve only a few factors; we found no case in which more than five factors were tested. In addition, although Plackett-Burman designs represent an important class of statistical models, we found no paper that applied a Plackett-Burman design to in-store testing. Finally, as we discuss in the last section of this paper, our methodology and results have useful implications for future research.

We organized the remainder of this paper as follows. In the *Literature Review* section, we review the relevant literature; in *The Supermarket Environment and the Factors Tested*, we describe and discuss the experimental environment and the factors tested. *Designing the Experiment* discusses the 24-run Plackett-Burman design that we implemented, compares its characteristics to those of two-level fractional designs, and describes our procedure for determining which stores to include and how to pair them. In the *Results* section, we discuss the results of the experiment, while in *Discussion of Results and Implications for Future Research*, we end the paper with a discussion of general conclusions and directions for future research.

Literature Review

Bisgaard (1992, p. 547) provides a notable, historical review of experimental design case studies that includes what he calls “a partial and unsystematic list of articles . . . showing engineering and manufacturing applications of experimental design.” This list comprises more than 130 case studies. More recent case studies applying experimental design methods to manufacturing problems are discussed by Lin and Chanada (2003), Cherfi et al. (2002), Schaub and Montgomery (1997), and Young (1996).

The literature on in-store testing includes research on product location and displays, advertising and promotions, pricing, and the effect of changes to the shopping environment. 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, whereas 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. Curhan (1974) used a two-level fractional factorial design to test the effects of price, newspaper advertising, display space, and display location on sales of fresh fruits and vegetables in supermarkets. In particular, he found that, for the four items tested, doubling display space increased sales from 28 to 49 percent. In a closely related study, Wilkinson et al. (1982) described a factorial experiment for assessing the impact of price, newspaper advertising, and display on the

sales of four products (bar soap, pie shells, apple juice, and rice) at a Piggly Wiggly grocery store. Their experiment considered three display levels (normal shelf space, expanded shelf space, and special display), and three price points (regular price, price cut, and deeper cut). Overall, the authors found large effects for expanded shelf space, very large effects for special display at the reduced price levels, and a large effect for special display even at the regular price (a sales increase of about 70 percent).

Many authors have studied the effect of changes in shelf-space position on sales. Drèze et al. (1994) used a basic test-control experimental approach to assess the sales impact of in-store shelf-space management. By changing the location of products among various shelf positions, they found that rearranging products in complementary groups and placing certain products at eye level could increase sales. Placing fabric softener between liquid and powder detergents, and moving toothbrushes from a top shelf to a shelf at eye level, increased category sales. They also found that shelf position was more important than the amount of shelf space allocated for a particular product. In earlier papers, several authors, including Brown and Tucker (1961) and Curhan (1972), studied shelf-space elasticities. Bultez and Naert (1988) and Bultez et al. (1989) studied space allocation using an attraction model to estimate brand interactions. Curhan (1972) found that across a large range of products shelf-space elasticity was approximately 0.2, meaning that a 40 percent increase in shelf space increased unit sales by 8 percent.

Several authors have studied the effect on sales of product displays. Chevalier (1975) used a factorial design in a supermarket setting to study the effect of special displays for 16 products in eight categories, with each display presented in conjunction with either a moderate or deep price cut. Gagnon and Osterhaus (1985) studied the effects of in-store display location and price on the sales of an antibiotic ointment.

In work related to shelf-space allocation and the effect of special displays, a number of authors, recognizing that more product on shelves and racks leads to more demand, have incorporated stock-level-dependent demand in inventory models. Urban (2005) provided a review of these models.

The work of Larson et al. (2005) and Hui et al. (2008) is particularly relevant to our work. Using RFID (radio frequency identification) tags attached to shopping carts, the authors were able to generate and analyze in-store shopping paths. Among their findings (Hui et al. 2008) are that shoppers spend a lot of time in the produce area, which is near the entrance in most stores, and that shoppers tend to backtrack after entering an aisle. In other words, rather than traversing through the aisle, a shopper is more likely to leave the aisle by retracing her steps. Hui et al. (2008) also mention that Sorensen (2003) observed that shoppers tend to move more quickly as they get closer to the checkout area. This suggests that we not place additional magazine locations near the checkout lane in our study, an insight that we followed in our experiment.

Woodside and Waddle (1975) used a Latin square design to study the effect an in-store promotion and a price cut on sales of instant coffee. Their point-of-sale advertising was a small, hand-lettered, five-by-seven-inch card with the brand name, size, price, the words “No Limit,” and a very small sticker with an image of the package. The authors deliberately made the advertising sign crude to make the effectiveness of this type of advertisement difficult to demonstrate. If they found it was effective in this case, it would strengthen the conclusion that this type of promotion would be effective in general. They found that the advertising card was most effective in conjunction with the price cut. However, they also found that, even without a price cut, sales increased by approximately 70 percent with the card.

The Supermarket Environment and the Factors Tested

The publisher owned a number of magazines. This experiment focused on the largest-selling weekly magazine in the portfolio—one of the top 10 circulated magazines in the United States. (For proprietary reasons, we are unable to reveal the name of the magazine.) Newsstand sales were an important source of revenue because the retail cover price was much higher than the per-issue price paid through annual subscriptions. Higher newsstand sales also increased advertising revenue because advertising fees increased with the number of magazines sold.

The publisher's marketing team decided to focus on one supermarket chain close to its corporate offices. This retailer had cooperated in previous tests, and many of the chain's supermarkets were located in the same market region, had strong magazine sales, and had a generally standard store layout. Because the marketing team considered the population and store design representative of the supermarket's national markets, the team believed that the test results would transfer well to other stores.

The digest-sized ($5\frac{1}{2} \times 7\frac{1}{2}$) weekly magazine had a relatively low cover price and was prominently placed in the checkout lanes to encourage impulse purchases. The magazine cover and content changed weekly and was customized by geographic region. All supermarkets in this test were located within the same region, yet far enough apart so that their customer bases did not overlap.

The project focused on the location, number, and arrangement of magazine racks as well as in-store advertising and promotions. Previous research described above showed that location and the amount of shelf space matters, and that simple in-store advertising can increase sales. Viewing sales of the magazine as mainly impulse purchases (Rook 1987), the operations team was particularly interested in the effect on sales of adding additional locations throughout the store. These added locations had been unused areas, and because the magazine displays required little space, testing additional locations was also of interest to the supermarket chain's management.

The work of Woodside and Waddle (1975), which we described earlier as showing the effectiveness of an on-shelf advertisement for coffee, seemed particularly relevant; the team identified a number of options for in-store advertising. For example, promotional messages could be added on to store shelves, magazine racks, grocery dividers, or the magazine itself.

After brainstorming a wide range of new ideas, the team identified 10 factors (Table 1). For each factor, the team selected two levels: the low (minus) level and the high (plus) level. A number of factors associated the number and location of "pockets" and "racks." A pocket is one slot that holds several copies of the same magazine in a magazine rack. A rack holds one or more pockets. Each pocket is like a wire cage with an open top and front; it holds about eight

Factor	(-) Low level	(+) High level
A Rack on cooler in produce aisle	No	Yes
B Location on checkout aisle	End cap	Over the belt
C Number of pockets on main racks	Current	More
D Rack by snack foods	No	Yes
E Advertise on grocery dividers	No	Yes
F Distribution of magazines in the store	Random	Even
G Oversized card insert	No	Yes, in 20% of copies
H Clip-on rack advertisement	No	Yes
I Discount on multiple copies	No	Yes
J On-shelf advertisement	No	Yes

Table 1: The team selected a low and high level for each of the 10 factors.

copies. The first copy faces forward with the others stacked behind. Increasing the number of pockets in a rack increases the magazine's exposure. The publisher pays the supermarket chain fees for each rack; these fees increase as the number of pockets increases. The team tested two new locations, one in the produce aisle and the other in the beer and snack food aisle. These choices were consistent with the findings of Hui et al. (2008) discussed above, who found that the produce aisle is a high-traffic area, and that shoppers appear to plan ahead and have target destinations in mind. The phenomenon of how they backtrack out of an aisle described above is also consistent with such behavior.

We briefly describe each factor and its levels below.

A: Rack on cooler in produce aisle. The team created a new, small rack with two pockets that was designed to fit on top of a refrigerated case in the center of the produce aisle. This rack, along with other test factors, was in addition to the main magazine rack locations. Each added rack increased the number of pockets by about 5 percent.

B: Location on checkout aisle. Two different rack locations were available at the checkout aisles: the end-cap racks that customers see as they approach the checkout and the over-the-belt racks above the moving grocery belt. The team had previously tried both locations but had never tested one against the other.

C: Number of pockets on main racks. With an incremental cost to each additional pocket, the marketing team wanted to estimate the effect of changing the number of pockets. Previous research described above

consistently showed that more shelf space leads to more sales. This experiment tested the current number of pockets versus an increase of 50 percent.

D: *Rack by snack foods.* The team developed a small rack to place on the shoulder-height snack food shelves; the aim was to entice shoppers who go directly to the beer/snack food aisles while on brief excursions.

E: *Advertise on grocery dividers.* With ever-growing alternatives for in-store advertising, from coupon dispensers to floor graphics and public address announcements, publishers have few low-cost options available. The operations team considered three new ideas (factors E, H, and J). One was to place short messages about the magazine on the four sides of the grocery dividers, the plastic sticks used to separate groceries at the checkout lane.

F: *Distribution of magazines in the store.* During the course of a week, sales across display racks will vary. For example, express lanes tend to have more customers, while some lanes that are open only during peak hours have fewer. This is a general phenomenon not limited to magazines: inventory levels of a given product show random variation across checkout lanes in store locations. Given this variation, the question is whether evening out the distribution of available inventory across locations would increase sales. We found no prior research that addressed this issue. The hypothesis was that a more even distribution of copies would increase sales. For all “F+” stores, the publisher paid an outside company to send “merchandisers” to each store midweek to even out the distribution of copies.

G: *Oversized card insert.* The one test factor that related to the magazine itself was a promotional card insert that was visible above the top of the magazine.

H: *Clip-on rack advertisement.* Small plastic clip-on signs with a promotional message were added to about half the racks in the checkout area.

I: *Discount on multiple copies.* Stickers on the front of each pocket offered a discount to customers who purchased two copies of the magazine.

J: *On-shelf advertisement.* The final in-store advertising factor was an on-shelf “billboard.” These small signs in plastic frames were attached to the edge of shelves and placed in a few specific product aisles.

Designing the Experiment

Fractional Designs, Confounding Among Main and Interaction Effects, and Relevant Design Issues

Factorial and fractional factorial designs are well-known. A factorial design tests all combinations of factors and levels; a “run” is a particular combination of factor settings. Because such a design is orthogonal, it provides independent estimates of all main effects, and all two-factor and higher-order interactions. Nearly all the factors in our study were binary in nature: display the magazine in an additional location or not, advertise on dividers or not, offer a promotion or not, and so forth. Consequently, we were able to employ a two-level design, with a design matrix consisting of -1 (low) and $+1$ (high) levels for the studied factors. A two-level factorial design with k factors consists of 2^k runs. However, in our case with 10 test factors, a factorial design would have required $2^{10} = 1,024$ experimental runs!

Fractional designs reduce the number of runs compared to full factorials; however, they introduce confounding among estimated effects and therefore uncertainty in the interpretation of results. The fewer the number of runs, the worse the confounding pattern. Two statistical principles are particularly useful in choosing among alternative designs (Box et al. 2005, Hamada and Wu 1992). First is the principle of “effects sparsity”: In any experiment the great majority of effects are likely to be negligible. Second is hierarchical ordering: Main effects tend to be larger in magnitude and hence more important than two-factor interactions, two-factor interactions larger than three-factor interactions, and so forth. In experiments for which four-factor and higher-order interactions can be estimated, they are almost certain to be negligible. In most cases, the three-factor interactions will be negligible as well.

Design resolution is an index that denotes the nature of the confounding in fractional designs. A resolution III design confounds main effects with two-factor interactions; a resolution IV design confounds main effects with three-factor interactions and confounds two-factor interactions with other two-factor interactions, whereas a resolution V design confounds main effects with four-factor interactions, and two-factor interactions with three-factor interactions.

We wanted a design that would give clear (unconfounded) estimates of all main effects—that is, a resolution IV design. With fractional factorial designs, the number of runs N is a power of 2 ($N = 4, 8, 16, 32$, and so forth). With 10 factors, a two-level fractional factorial design requires a minimum of 16 runs. A 16-run fractional factorial design is resolution III with main effects confounded with two-factor interactions. One can construct a fractional factorial design in 32 runs that is resolution IV (Box et al. 2005); however, it would double the number of runs compared to the 16-run resolution III design. Because at least one store is needed for each experimental run, a minimum of 32 stores would be needed. In addition, as we discuss below, it was advantageous to have two stores for each experimental run to reduce the variance of the experimental error. Thus, 64 stores would be needed, a number that the supermarket firm viewed as excessive, given the cost of setting up and monitoring the experiment. As we describe in the next section, this led us to consider the class of Plackett-Burman designs.

Plackett-Burman Designs

In a classic paper, Plackett and Burman (1946) showed how to construct two-level orthogonal designs when the number of runs N is a multiple of 4 ($N = 4, 8, 12, 16, 20, 24$, and so on). If the run size is a power of 2, these designs are identical to two-level fractional factorial designs. Because the number of runs N in two-level fractional designs is a power of 2 ($N = 4, 8, 16, 32, 64$, and so forth), these designs leave large gaps in the run sizes; the Plackett-Burman designs fill in these gaps. Table 2 shows the 12-run Plackett-Burman design.

Constructing the 24-Run Plackett-Burman Design Used in the Experiment

Table 3 shows the 24-run design matrix we used in our experiment. The first 12 rows consist of the 12-run Plackett-Burman design of resolution III shown in Table 2 with main effects confounded with two-factor interactions. We augmented this design by adding 12 additional runs employing a “foldover.” As discussed in the appendix, the resulting design has resolution IV with main effects no longer confounded with two-factor interactions. We have switched the signs in all columns for the foldover. The term “reflection” is sometimes used to describe the 12 additional

Run	Factors										
	A	B	C	D	E	F	G	H	I	J	K
1	+	+	-	+	+	+	-	-	-	+	-
2	+	-	+	+	+	-	-	-	+	-	+
3	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	-	-	-	+	-	+	+	-
5	+	+	-	-	-	+	-	+	+	-	+
6	+	-	-	-	+	-	+	+	-	+	+
7	-	-	-	+	-	+	+	-	+	+	+
8	-	-	+	-	+	+	-	+	+	+	-
9	-	+	-	+	+	-	+	+	+	-	-
10	+	-	+	+	-	+	+	+	-	-	-
11	-	+	+	-	+	+	+	-	-	-	+
12	-	-	-	-	-	-	-	-	-	-	-

Table 2: This table shows the 12-run Plackett-Burman design that can be used to study main effects of up to 11 factors.

runs with every “+” and “-” switched, somewhat like holding a mirror up to the original design. For example, run 1 has A+, B+, C-, D+, E+, F+, G-, H-, I-, J+, K-. For the first reflected run (run 13), all signs are reversed to become A-, B-, C+, D-, E-, F-, G+, H+, I+, J-, K+.

Selection and Pairing of Test Stores

We needed a minimum of 24 stores with our 24-run design, one for each experimental run (combination of factor settings). However, it is beneficial to have two stores at each experimental run because the response (percentage change in sales) of a two-store unit is less variable than the response of a single store. By reducing the variation, two stores per run allows us to establish statistical significance more easily; i.e., it increases the statistical power of our test. The 48 stores in our design were less than the 50 stores that the supermarket chain specified as the maximum number for the test.

If stores vary widely in sales levels, then they should not be grouped together in the same run. This is because our confidence in a 10 percent change in sales for a store that sells 10 magazines one week and 11 the next is very different from our confidence for a store that sells 100 magazines one week and 110 the next. Hence, it is advantageous to match smaller stores with larger stores so that the average sales volume per run is relatively constant. After grouping stores by size, we randomly paired larger (higher-demand) stores with smaller stores so that

Run	Rack on cooler in produce aisle	Location on checkout aisle	Number of pockets on main racks	Rack by snack foods	Advertise on grocery dividers	Distribution of magazines in the store	Oversized card insert	Clip-on rack advertisement	Discount on multiple copies	On-shelf advertisement	(Empty)
	A	B	C	D	E	F	G	H	I	J	K
1	+	+	-	+	+	+	-	-	-	+	-
2	+	-	+	+	+	-	-	-	+	-	+
3	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	-	-	-	+	-	+	+	-
5	+	+	-	-	-	+	-	+	+	-	+
6	+	-	-	+	+	-	+	+	-	+	+
7	-	-	-	-	-	+	+	-	+	+	+
8	-	-	+	-	+	+	-	+	+	+	-
9	-	+	-	+	+	-	+	+	+	-	-
10	+	-	+	+	-	+	+	+	-	-	-
11	-	+	+	-	+	+	+	-	-	-	+
12	-	-	-	-	-	-	-	-	-	-	-
13	-	-	+	-	-	-	+	+	+	-	+
14	-	+	-	-	-	+	+	+	-	+	-
15	+	-	-	-	+	+	+	-	+	-	-
16	-	-	-	+	+	+	-	+	-	-	+
17	-	-	+	+	+	-	+	-	-	+	-
18	-	+	+	+	-	+	-	-	+	-	-
19	+	+	+	-	+	-	-	+	-	-	-
20	+	+	-	+	-	-	+	-	-	-	+
21	+	-	+	-	-	+	-	-	-	+	+
22	-	+	-	-	+	-	-	-	+	+	+
23	+	-	-	+	-	-	-	+	+	+	-
24	+	+	+	+	+	+	+	+	+	+	+

Table 3: The reflected Plackett-Burman design in 24 runs is shown.

the combined mean weekly demands for the 24 pairs were relatively constant. This pairing helped stabilize the variability across the 24 test runs.

We used the weekly average number of copies sold over the seven previous weeks for each pair of stores as the baseline for the test, which we ran for two weeks. Table 4 shows the raw data and percent changes in sales for each of the 24 runs. For example, consider the data for run 1. The baseline—average sales over the previous seven weeks—was 97.0 for the pair of stores (average 69.8 for store 1 plus average 27.2 for store 2). During week 1 of the experiment, sales totaled 115 (79 copies in store 1 and 36 in store 2). The percentage change for sales in week 1 was $(115 - 97)/97 = 0.186$ or 18.6, while the percentage change in week 2 was 29.9. The average of the week 1 and week 2 percentage changes, which is our response variable, was 24.23.

The 48 stores included in the experiment were chosen from the nearly 100 stores in the region. The authors reviewed the most recent six months of data

for all stores. Control charts (individuals (X) and moving range (MR) charts) of weekly sales allowed us to identify stores that had statistically stable (in control) weekly sales and to eliminate stores for which the control charts showed special causes of variation. Montgomery (2004) included a discussion of control charts for individual measurements. New stores and stores with dramatic changes of growth were eliminated from consideration, as were stores with low sales volumes. We also eliminated stores near colleges, resort areas, or in other locations in which sales changed dramatically during the year. Of the remaining stores, we selected the 48 with the highest weekly sales for the experiment. All stores were in the same region, but far enough apart to have different customer bases (previous analyses by the retailer determined that few people regularly shopped in more than one of the experimental locations). All supermarkets had similar store layouts and comparable numbers of magazine racks. The number of pockets the publisher paid to fill and the number of copies of the magazine

Run	Number of												Baseline unit sales				Key metric:								
	Rack on produce aisle cooler in aisle	Location on checkout aisle	Aisle	B	C	D	E	F	G	H	I	J	K	Store 1	Store 2	Store 1+2	Week 1 units sold	Week 2 units sold	Week 1	Week 2	Average	% change in sales per 2-store test unit/week (actual – baseline)/baseline (%)			
1	+	+	+	+	+	+	+	+	+	+	+	+	+	69.8	27.2	97.0	79	36	115	88	38	126	18.6	29.9	24.23
2	+	+	+	+	+	+	+	+	+	+	+	+	+	42.8	63.3	106.1	47	79	126	52	71	123	18.8	15.9	17.34
3	+	+	+	+	+	+	+	+	+	+	+	+	+	39.0	66.0	105.0	37	58	95	44	61	105	-9.5	0.0	-4.76
4	+	+	+	+	+	+	+	+	+	+	+	+	+	54.0	81.8	135.8	69	111	180	60	74	134	32.5	-1.3	15.61
5	+	+	+	+	+	+	+	+	+	+	+	+	+	103.7	45.7	149.4	97	42	139	105	36	141	-7.0	-5.6	-6.29
6	+	+	+	+	+	+	+	+	+	+	+	+	+	58.8	102.8	161.6	79	110	189	72	109	181	17.0	12.0	14.48
7	+	+	+	+	+	+	+	+	+	+	+	+	+	52.3	56.5	108.8	45	58	103	47	60	107	-5.3	-1.7	-3.49
8	+	+	+	+	+	+	+	+	+	+	+	+	+	76.3	65.2	141.5	103	48	151	85	39	124	6.7	-12.4	-2.83
9	+	+	+	+	+	+	+	+	+	+	+	+	+	50.2	57.8	108.0	63	54	117	51	56	107	8.3	-0.9	3.70
10	+	+	+	+	+	+	+	+	+	+	+	+	+	59.8	53.3	113.1	60	46	106	79	64	143	-6.3	26.4	10.08
11	+	+	+	+	+	+	+	+	+	+	+	+	+	95.3	40.8	136.1	82	44	126	78	40	118	-7.4	-13.3	-10.36
12	+	+	+	+	+	+	+	+	+	+	+	+	+	40.3	83.0	123.3	34	77	111	34	73	107	-10.0	-13.2	-11.60
13	+	+	+	+	+	+	+	+	+	+	+	+	+	79.0	74.2	153.2	94	86	180	94	72	166	17.5	8.4	12.92
14	+	+	+	+	+	+	+	+	+	+	+	+	+	63.7	52.0	115.7	51	46	97	62	56	118	-16.2	2.0	-7.09
15	+	+	+	+	+	+	+	+	+	+	+	+	+	45.7	64.8	110.5	44	71	115	41	59	100	4.1	-9.5	-2.71
16	+	+	+	+	+	+	+	+	+	+	+	+	+	88.3	63.0	151.3	119	73	192	101	65	166	26.9	9.7	18.31
17	+	+	+	+	+	+	+	+	+	+	+	+	+	52.4	95.0	147.4	59	110	169	67	87	154	14.7	4.5	9.57
18	+	+	+	+	+	+	+	+	+	+	+	+	+	102.2	33.2	135.4	96	28	124	98	28	126	-8.4	-6.9	-7.68
19	+	+	+	+	+	+	+	+	+	+	+	+	+	63.0	50.5	113.5	66	62	128	69	66	135	12.8	18.9	15.86
20	+	+	+	+	+	+	+	+	+	+	+	+	+	45.5	68.7	114.2	68	69	137	69	79	148	20.0	29.6	24.78
21	+	+	+	+	+	+	+	+	+	+	+	+	+	26.5	98.8	125.3	30	107	137	24	104	128	9.3	2.2	5.75
22	+	+	+	+	+	+	+	+	+	+	+	+	+	59.8	46.2	106.0	52	64	116	63	61	124	9.4	17.0	13.21
23	+	+	+	+	+	+	+	+	+	+	+	+	+	79.7	59.2	138.9	89	63	152	99	57	156	9.4	12.3	10.87
24	+	+	+	+	+	+	+	+	+	+	+	+	+	76.0	45.0	121.0	72	48	120	71	50	121	-0.8	0.0	-0.41

Table 4: The test results of the reflected Plackett-Burman design in 24 runs are shown.

delivered to each store varied based on historical sales levels. The publisher had a proprietary model to calculate the number of copies (called “draw”) delivered to each store every week. For the experiment, the draw was increased by a set percentage across all supermarkets to ensure all pockets had a reasonable number of copies at the start of each test week.

Selection bias is a possibility in the nonrandom choice of the 48 experimental units. Certain stores in the supermarket chain could show unique effects apart from the selected group of “in-control” stores. High-variability, unstable stores could provide insights into possibly unique factors that influence behavior in seasonal, high-growth, or other market environments. These “special cause” stores also add to the experimental error and potential for outliers within the test data, however. Focusing on the most stable and representative stores for the initial test increases the opportunity to identify a larger number of more reliable effects.

As noted by Wilkinson et al. (1982), Hirschman and Stampfl (1980) considered retail stores to be a natural laboratory for understanding consumer behavior because experiments in actual settings reduce threats to external validity. With 48 stores and 10 factors in our test, it was especially important to ensure that each store had the right experimental setup and that it was maintained over the two-week test period. Two managers at each store were designated as coordinators, and at least one worked during all hours of the test period. The publisher’s team members trained the coordinators, communicated frequently with them, and visited each store at least once each week of the test.

Results

Estimating Main Effects and Determining Statistical Significance

The main effects are obtained by applying the + and – signs in the design columns to the responses (averages) in the last column of Table 4, and dividing the resulting sum by 12 (the number of plus signs). Alternatively, one can regress the averages on the design vectors. The only difference is that the regression coefficients are one-half the estimated main effects.

We treat the changes in weeks 1 and 2 as independent replications; we calculate an estimate of the variance of individual measurements. (We confirmed the absence of serial correlation among weekly percentage changes by calculating, for each run separately, the difference between successive changes. We found that exactly half of these differences were positive and half were negative, making serial correlation unlikely.) For example, for the first run the variance estimate is $[(18.6 - 24.23)^2 + (29.9 - 24.23)^2]/1 = 64.3$. We pool the 24 variances to obtain the overall estimate $s^2 = 88.047$; it measures the experimental error among weekly changes. The variance of the average for weeks 1 and 2 that goes into the main effects calculation is given by $s^2/2$. Each effect is the difference of two averages of 12 responses; hence, $\text{Var}(\text{effect}) = (s^2/2 + s^2/2)/12 = s^2/12 = 88.047/12 = 7.34$; the standard error of an effect is given by $\sqrt{7.34} = 2.71$. Effects that are larger than 1.96 times the standard error (that is, larger than 5.31) are considered significant. Figure 1 displays the effects graphically. Three effects (A, F, and D) were statistically significant; factor E was not quite statistically significant. We briefly describe each below.

A+: *Rack on cooler in produce aisle.* The display on top of the refrigerated case in the produce section increased sales by 9.97 percent. This identified the one most profitable new location to sell the magazines and supported the idea of placing the display in the high-traffic produce area. This was a major change, placing magazines far from their usual location. The rack was also a creative new design that took up little space in a previously unused location.

F–: *Not adjusting the number of magazines among pockets.* Sales dropped 8.71 percent when merchandisers

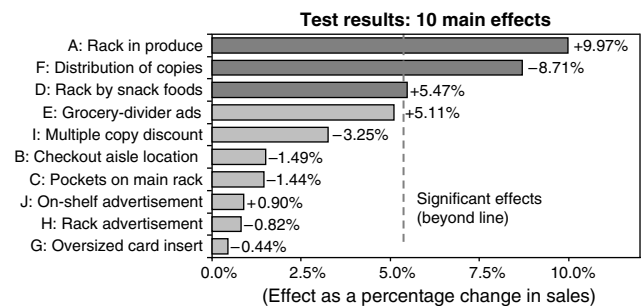


Figure 1: The graph shows the estimated effects.

adjusted the number of magazines among the pockets. Although this result was surprising, it saved a great deal of money that might have been wasted on unnecessary effort in the future. An explanation for these results is that empty slots imply scarcity, infer high demand, and increase impulse buying. It is also possible that there was no actual difference and the observed difference occurred by chance. In either case, there was no evidence to support redistributing the magazines among the pockets.

D+: *Rack by snack foods*. The second new display with a large effect was a magazine rack at the other end of the store next to snack foods and beer, increasing sales by 5.47 percent. Once again, a small rack in a completely new location was beneficial.

E+: *Grocery-divider ads*. This large, although narrowly statistically significant, effect related to the advertising on the grocery dividers at the checkout counters. With its low cost and the high interest of the marketing team, the publisher decided to pursue this idea in the future.

The racks of the two significant rack-related factors (factors A and D) held the same number of magazines but were of different shapes and dimensions; each factor tested the combination of two elements: location and design of the rack. A follow-up experiment that considers location and design as two separate factors could further determine the effects of rack location and rack design.

Profitability analyses (not shown here), which considered the cost of the new racks and the cost of the divider ads, showed that the projected increased sales were much higher than the associated costs.

Learning that some effects were nonsignificant was also useful. Except for the grocery-divider ads, in-store advertising showed no impact; therefore, the publisher discarded the on-shelf and clip-on ads. The location of racks on the checkout aisle was not statistically significant. Because the estimated effect was negative, however, the team decided to keep using the racks over the belt. The second-copy discount and oversized-card insert had no impact and were eliminated. Surprisingly, more pockets on the main racks, which is analogous to increasing a product's shelf space, had no impact on sales.

As we discuss in the appendix, one-half the difference of the main effects obtained from the first set of 12 runs and from its foldover provides an estimate of the weighted sum of the two-factor interactions that prior to the foldover were confounded with that main effect. We performed these calculations for each of the 10 factors and found that each of these weighted sums of two-factor interactions was small in magnitude and not close to being statistically significant.

Further Comments

In our experiment, which ran for two weeks, $\text{Var}(\text{effect}) = (s^2/2 + s^2/2)/12 = s^2/12$. In general, if n is the number of weeks in the experiment and if we can assume independence of changes from week to week, $\text{Var}(\text{effect}) = (s^2/n + s^2/n)/12 = s^2/6n$. An increase in the number of weeks, n , would have reduced the variance of an effect and increased the power of the test.

We initially recommended running the test for five weeks. We obtained an estimate of the variability of weekly changes by pooling the information from the control charts across all 24 pairs of stores. We found that, on average, 125 copies were sold each week, with a standard deviation of 12 copies or about 10 percent. Power calculations using the Minitab software (with a 10 percent standard deviation and a 5 percent significance level) showed that a five-week test would have an 80 percent chance of detecting an effect that impacted sales by 5 percent. However, as the launch date approached, publishing company executives felt the need to speed up the project and reduce costs; therefore, they limited the test to only four weeks. Then, just before the test began, further delays reduced its length to two weeks.

Testing all factors simultaneously has very large sample-size advantages compared to testing each of the 10 factors one at a time. Suppose the publisher had decided to start with the test of a single change, such as adding a display rack in the produce section. To obtain the same degree of precision in estimating the effect of this change as we achieved in our experiment, 12 pairs of stores would need to implement the change for two weeks, while another 12 pairs would be used for comparison purposes. Testing the nine other factors in this manner would require an

additional 18 weeks. Because holiday periods must be avoided, the total of 20 weeks would take place over an even longer period.

There is one additional, important advantage in testing factors simultaneously—namely, the ability to shed light on the presence and identity of two-factor interactions. In contrast, the approach of changing one factor at a time cannot identify interactions; if such interactions are present, this approach could lead to incorrect conclusions. In this experiment, we were able to obtain clear estimates of main effects and show that interaction terms were negligible. An omnibus test for the absence of interactions can be constructed by applying a regression lack-of-fit test (see, for example, Abraham and Ledolter 2006, pp. 199–204). Treating the weeks 1 and 2 percentage changes in Table 4 as replicate responses at the 24 different experimental conditions leads to the pure error sum of squares of 2,113.1, with 24 degrees of freedom (which, by the way, results in our variance estimate $s^2 = 2,113.1/24 = 88.047$ in the *Estimating Main Effects and Determining Statistical Significance* section). Fitting the main-effects model with 10 factors leads to the error sum of squares of 4,883.6, with $48 - 10 - 1 = 37$ degrees of freedom. The difference in these two sums of squares represents the lack of fit of the main-effects model. The F -statistic for lack of fit, $((4,883.6 - 2,113.1)/(37 - 24))/(2,113.1/24) = 1.85$ with probability value 0.093, is insignificant at the 0.05 significance level.

Discussion of Results and Implications for Future Research

With a two-week test of 10 variables, placing a new rack in the produce area was the biggest winner, with the addition of a rack in the snack foods aisle showing a significant benefit. The team avoided unnecessary operating costs when four of the five in-store advertising factors showed no evidence of a positive effect. The common perception that redistributing copies was a worthwhile investment proved to be a significant misconception. This surprising result is worthy of future research, including experiments in other retail settings.

Our results showing the benefits of additional display locations are consistent with previous research

on special displays, although the magnitude of the effects we found are much smaller than those found in previous studies. This is not surprising because the visual impact of a display rack of magazines cannot be compared to that of a very large end-of-aisle display of packaged goods. A surprising result related to adding pockets on the main magazine racks: Previous research clearly showed that increasing shelf space leads to higher sales. However, our test showed no sales benefit although the number of display pockets was increased by 50 percent. This outcome needs further study to reconcile these conflicting results. With respect to in-store advertising and promotions, the only changes that showed some evidence of effectiveness were the ads on the grocery dividers. However, given that in-store promotions are inexpensive and easy to implement, their effectiveness warrants further study. Moreover, there are indications that the use of in-store advertising is likely to increase dramatically in the near future. According to Herb Sorenson, a leading consultant to the supermarket industry, “There will be a huge growth in the use of in-store media to try to influence the way shoppers navigate a store and what they buy... \$300 billion of advertising money will move into the retail space in the next five years” (Knowledge@Wharton 2006). Plackett-Burman and fractional factorial experiments that enable rapid testing of content and in-store locations will be useful in determining the best choices.

Our work complements the work of Hui et al. (2008) on shopping paths. Using path data and the insights that are likely to derive from these authors’ modeling efforts, it should be possible to identify interesting hypotheses and potentially attractive locations for testing using our multivariable methodology.

In this paper, we have emphasized the importance of fractional experimental designs in general, and Plackett-Burman designs in particular. Although previous research has used experimental design methods in retail testing, these studies have typically utilized factorial designs that test all combinations of factor levels and have involved relatively few factors. In contrast, fractional designs, such as the one we used in our study, dramatically reduce the number of experimental runs required and make it possible

to test many factors simultaneously. Although their application to manufacturing problems has been widespread, the use of fractional designs in service contexts has been rare.

Fractional designs allow the investigator to study the effects of more factors in fewer runs. Being able to learn about the effects of several factors in a fraction of the time required by one-factor tests speeds up the learning curve and facilitates an earlier implementation of the winning combination of factors. This time advantage and the cost savings that result from the reduced number of runs make these designs important.

Our purpose in this paper has been both to report on a successful application in retail testing and to highlight the opportunities that arise if Plackett-Burman and fractional factorial methods are applied to service problems. In general, multifactor experiments in service operations can be used to study the effects on service quality and performance of changes in staffing, training levels, procedures, and service-system design. Specific examples include the design of websites; direct mail campaigns for magazines, credit cards, as well as other products; and, as this paper illustrates, various in-store experiments to evaluate changes in factors, such as package design, price, and point-of-sale displays. Website design is a particularly attractive area for multifactor experiments. Firms traditionally test one change to a “landing page” at a time. In contrast, fractional designs offer the opportunity to test 10 to 20 changes over a few weeks, compared to the months or even years that would be required by testing one factor at a time. Chapter 8 of Ledolter and Swersey (2007) provides a detailed website-design case study. Similarly, direct mail is ideally suited for simultaneous testing of multiple factors using fractional designs (Bell et al. 2006, Ledolter and Swersey 2006). The traditional approach to direct mail testing has been to test one change in a particular mailing—so called “champion/challenger” testing. However, multifactor testing using fractional designs has major advantages in terms of sample-size reduction and the ability to estimate interactions. As researchers become more aware of these powerful and efficient approaches, their use is likely to become more widespread.

Appendix

Confounding in Plackett-Burman Designs and the Construction of Resolution IV Designs Through Foldovers

Compared to fractional factorial designs, Plackett-Burman designs have more-complex confounding patterns. In the 12-run design in Table 2, the main effect of each factor is confounded with all two-factor interactions involving the other factors. However, in contrast to fractional factorial designs, 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 with correlation coefficient $|\rho| < 1$. Therefore, it can be shown (Ledolter and Swersey 2007, Chapter 6) that estimating the main effect of a particular factor, by taking the difference in the response between the high (plus) and low (minus) levels for that factor, 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. Barrentine (1996) includes 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 12-run Plackett-Burman design in Table 2, the weights (correlations) are either $-1/3$ or $+1/3$. For example, consider the main effect of factor A and the BE interaction. We use $+1$ and -1 to represent the column signs and multiply the entries in columns B and E to obtain the entries in column BE. 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
Column A	+1	+1	-1	+1	+1	+1	-1	-1	-1	+1	-1	-1
Column BE	+1	-1	-1	-1	-1	-1	+1	-1	+1	+1	+1	+1

Both columns have six plus and six minus signs; the entries in each column add to zero. Furthermore, the sum of the squares of the entries in each column is 12, the number of runs N . The columns are correlated. In 4 of the 12 runs the signs match, while the signs

are opposite in 8 runs. The correlation between these two mean zero columns (call them x and z) is given by

$$\rho = \frac{\sum x_i z_i}{\sqrt{\sum x_i^2} \sqrt{\sum z_i^2}} = \frac{-4}{12} = -1/3.$$

The correlation of $-1/3$ indicates that there is some linkage between the signs of these columns; however, the correlation is rather weak.

These correlations determine the confounding. Consider factor A , for example, and its column of signs. Let l_A denote the average of the responses when factor A is at its $+$ level minus the average of the responses when factor A is at its minus level; it is also called a contrast because it compares two averages. Let A represent the main effect of A , BC the interaction between factors B and C , BD the interaction between factors B and D , and so on. Then, the contrast l_A for column A estimates

$$\begin{aligned} & A + \frac{1}{3}(BF + BI + BJ + CD + CG + CI + DE + DF + EJ \\ & \quad + EK + FH + GH + GJ + HK + IK) \\ & - \frac{1}{3}(BC + BD + BE + BG + BH + BK + CE + CF \\ & \quad + CH + CJ + CK + DG + DH + DI + DJ + DK \\ & \quad + EF + EG + EH + EI + FG + FI + FJ + FK \\ & \quad + GI + GK + HI + HJ + IJ + JK). \end{aligned}$$

The main effect of factor A is confounded with 45 two-factor interactions, with weight $1/3$ for 15 of the interactions and weight $-1/3$ for the other 30. To show the consequences of the confounding in the 16-run fractional factorial design and the 12-run Plackett-Burman design (Table 2), assume there is a single statistically significant two-factor interaction. Although this one interaction will bias 9 of the 11 main effects in the Plackett-Burman design and at most one main effect in the fractional factorial design, the Plackett-Burman design has an advantage. Because the correlations are all plus or minus $1/3$, the magnitude of the bias will be only one-third of the magnitude of the interaction. In contrast, for the fractional factorial design, the single main effect will be biased by the full magnitude of the two-factor interaction.

Although we would have expected the main-effect biases to be rather small if we had used the 12-run Plackett-Burman design, we chose to eliminate them

completely by creating a resolution IV design in 24 runs. We augmented the 12-run Plackett-Burman design in Table 2 by adding the 12 additional runs in Table 3 and by employing a foldover that switches the signs in all columns. As we show below, the foldover of the Plackett-Burman design leads to a resolution IV design in which main effects are no longer confounded with two-factor interactions.

We saw earlier that the contrast l_A from the original 12 runs is an estimate of $A + (1/3)T_1 - (1/3)T_2$, where T_1 and T_2 represent the interactions in the first and second bracketed terms. For the 12 foldover runs, let l_A^f denote the average of the responses when factor A is at its plus level minus the average of the responses when factor A is at its minus level. Because the sign of every column is switched, the signs for the two-factor interaction columns are unchanged. However, because the signs for column A have been switched, the correlations between A and the two-factor interactions not involving A have their signs changed. Thus, l_A^f estimates $A - (1/3)T_1 + (1/3)T_2$. We use the two simple algebraic operations of addition and subtraction to combine the estimated effects from the original 12 runs and the additional 12 runs of the foldover, and to reveal the confounding pattern for the entire 24-run experiment. One-half of their sum $(l_A + l_A^f)/2$ estimates A (the main effect of A), while one-half of their difference $(l_A - l_A^f)/2$ estimates $(1/3)T_1 - (1/3)T_2$ (the weighted sum of the two-factor interaction terms). These two operations separate A (the main effect of factor A) from the weighted sum of the interaction terms, and provide estimates of each. Repeating this procedure for the other 10 factors (including the unused column K) provides for each factor a clear estimate of its main effect and an estimate of the weighted sum of its previously confounded two-factor interactions.

Acknowledgments

We gratefully acknowledge the constructive comments of the editor, an associate editor, and two referees.

References

- Abraham, B., J. Ledolter. 2006. *Introduction to Regression Modeling*. Thomson, Brooks/Cole, Belmont, CA.
- Barclay, W. D. 1969. Factorial design in a pricing experiment. *J. Marketing Res.* 6(4) 427–429.

- Barrentine, L. B. 1996. Illustration of confounding in Plackett-Burman designs. *Quality Engrg.* 9(1) 11–20.
- Bell, G. H., J. Ledolter, A. J. Swersey. 2006. Experimental design on the front lines of marketing: Testing new ideas to increase direct mail sales. *Internat. J. Res. Marketing* 23(3) 309–319.
- Bisgaard, S. 1992. Industrial use of statistically designed experiments: Case study references and some historical anecdotes. *Quality Engrg.* 4(4) 547–562.
- Box, G. E. P., W. G. Hunter, J. S. Hunter. 2005. *Statistics for Experimenters*, 2nd ed. John Wiley & Sons, New York. (orig. pub. 1978.)
- Brown, W., W. T. Tucker. 1961. The marketing center: Vanishing shelf space. *Atlanta Econom. Rev.* 11(October) 9–13.
- Bultez, A., P. Naert. 1988. S.H.A.R.P.: Shelf allocation for retailer's profit. *Marketing Sci.* 7(3) 211–231.
- Bultez, A., E. Gijbrecchts, P. Naert, P. Vanden Abeele. 1989. Asymmetric cannibalism in retail assortments. *J. Retailing* 65(2) 153–192.
- Cherfi, Z., B. Bechard, N. Boudaoud. 2002. Case study: Color control in the automotive industry. *Quality Engrg.* 15(1) 161–170.
- Chevalier, M. 1975. Increase in sales due to in-store display. *J. Marketing Res.* 12(4) 426–431.
- Curhan, R. C. 1972. The relationship between shelf space and unit sales in supermarkets. *J. Marketing Res.* 9(4) 406–412.
- Curhan, R. C. 1974. The effects of merchandising and temporary promotional activities on the sales of fresh fruits and vegetables in supermarkets. *J. Marketing Res.* 11(3) 286–294.
- Drèze, X., S. J. Hoch, M. E. Purk. 1994. Shelf management and space elasticity. *J. Retailing* 70(4) 301–326.
- Fisher, R. A. 1966. *The Design of Experiments*, 2nd ed. Hafner Publishing Company, New York. (orig. pub. 1935.)
- Gagnon, J. P., J. T. Osterhaus. 1985. Effectiveness of floor displays on the sale of retail products. *J. Retailing* 61(Spring) 104–116.
- Hamada, M., C. F. J. Wu. 1992. Analysis of designed experiments with complex aliasing. *J. Quality Tech.* 24(3) 130–137.
- Hirschman, E., R. Stampfl. 1980. Retail research: Problems, potentials and priorities. E. Hirschman, R. Stampfl, eds. *Competitive Structure in Retail Markets*. American Marketing Association, Chicago, 68–77.
- Holland, C. W., D. W. Cravens. 1973. Fractional factorial designs in marketing research. *J. Marketing Res.* 10(3) 270–276.
- Hui, S. K., P. S. Fader, E. T. Bradlow. 2008. Path data in marketing: An integrative framework and prospectus for model building. *Marketing Sci.*, ePub ahead of print October 9, <http://mktsci.journal.informs.org/cgi/content/abstract/mksc.1080.0400v1>.
- Knowledge@Wharton. 2006. The “traveling salesman” goes shopping: The efficiency of purchasing patterns in the grocery store. Retrieved November 23, 2008, <http://knowledge.wharton.upenn.edu/article.cfm?articleid=1608>.
- Larson, J. S., E. T. Bradlow, P. S. Fader. 2005. An exploratory look at supermarket shopping paths. *Internat. J. Res. Marketing* 22(4) 395–414.
- Ledolter, J., A. J. Swersey. 2006. Using a fractional factorial design to increase direct mail response at *Mother Jones Magazine*. *Quality Engrg.* 18(4) 469–475.
- Ledolter, J., A. J. Swersey. 2007. *Testing 1–2–3: Experimental Design with Applications in Marketing and Service Operations*. Stanford University Press, Stanford, CA.
- Lin, T., B. Chanada. 2003. Quality improvement of an injection-molded product using design of experiments: A case study. *Quality Engrg.* 16(1) 99–104.
- Montgomery, D. C. 2004. *Introduction to Statistical Process Control*, 5th ed. John Wiley & Sons, New York.
- Plackett, R. L., J. P. Burman. 1946. The design of optimum multifactorial experiments. *Biometrika* 33(4) 305–325.
- Rook, D. W. 1987. The buying impulse. *J. Consumer Res.* 14(2) 189–199.
- Schaub, D. A., D. C. Montgomery. 1997. Using experimental design to optimize the stereolithography process. *Quality Engrg.* 9(4) 575–585.
- Sorensen, H. 2003. The science of shopping. *Marketing Res.* 15(3) 30–35.
- Urban, T. L. 2005. Inventory models with inventory-level-dependent demand: A comprehensive review and unifying theory. *Eur. J. Oper. Res.* 162(3) 792–804.
- Wilkinson, J. B., J. B. Mason, C. H. Paksoy. 1982. Assessing the impact of short-term supermarket strategy variables. *J. Marketing Res.* 19(1) 72–86.
- Woodside, G., G. L. Waddle. 1975. Sales effects of in-store advertising and price specials. *J. Advertising Res.* 15(June) 29–34.
- Young, J. C. 1996. Blocking, replication, and randomization—The key to effective experimentation: A case study. *Quality Engrg.* 9(2) 269–277.

Editorial note: The company wishes to remain anonymous.

Joni L. Keith, Statistician, Bush Brothers & Company, 1016 East Weisgarber Road, Knoxville, Tennessee 37909-2683, writes: “I am writing to verify the experimental design and validity of the results from the market test presented in the article, ‘A Plackett-Burman Experiment to Increase Supermarket Sales of a National Magazine.’

“My statistical work with the publisher X led to my involvement with this project with one of the authors, Gordon Bell (since that time, all of us involved in the project are no longer working with the company).

“This project was very successful and insightful. All previous experiments were simple champion/challenger tests that would have never allowed X to test this large a number of variables or uncovered such accurate insights. This work really opened management’s eyes to the value of multivariable testing.

“The experiment was conducted as described and the results were statistically valid with a clear market impact.”