



## Case Study: Office Supplies E-mail Test

### *Fast test, small market, big results*

*This case is also presented in a new book by Johannes Ledolter and Arthur Swersey, professors at the University of Iowa and Yale School of Management, Design of Experiments: Statistical Methods for Improving Quality and Performance*

### Introduction

Real-world marketing tests always have some challenges. But one office-supplies retailer had more than their share trying to improve e-mail ROI. With few e-mail addresses, one campaign a month, and an inexperienced team, they had trouble improving performance.

This all changed with one test. In one drop, they tested 12 changes to the e-mail creatives and offers across three customer segments with just 34,060 names. Test results identified four significant effects and one significant interaction for a 54% jump in response and an even larger increase in profit.

This case study is statistically complex (with details footnoted), yet was surprisingly simple for the marketing team—they needed just a few additional days to plan, create, and launch the 32-recipe test. Note that the detailed explanation of the statistics offers deeper insight into what goes on “behind the scenes,” but for the marketing team, the test was straightforward and the benefits were clear...

- **They tested 15 variables as quickly as they could test one.**
- **They learned in one week what would have taken 8 months with split-run tests.**
- **Results quantified four significant elements and one important two-way interaction.**

### Planning the Test

With fast response, low costs, and flexible production, e-mail was a great place to start testing. In addition, what worked in e-mail could then be tested in the catalog or even in retail stores.

However, in this early stage of their business-to-business e-mail program, the Internet marketing director had a list of only 35,000 e-mail addresses, covering three distinct customers segments each with different buying behavior. With so few names, they could test two or three different versions of the e-mail in each monthly drop, but even these seldom led to much improvement.

A consultant was brought on to help. He agreed with the Internet director that sample size could be a problem. In this case, with only 35,000 names and an average response rate of 1%, an effect would have to change response by about 20% (for example, from 1.0% to 1.2%) to be significant<sup>1</sup>. Even with these constraints, the consultant offered encouragement that testing could be very profitable.

## Test Elements

After brainstorming and trimming the list down to the boldest ideas, the marketing team identified 13 variables and selected two different versions of each variable to test. These 13 elements could be tested simultaneously in a 16-recipe test design. However, in this case, the consultant recommended a 32-recipe “fractional-factorial” test design, so he could clearly analyze a number of potential interactions<sup>2</sup>.

The three customer segments also had to be considered, but a three-level test element would lead to an unbalanced design. Instead, the three segments were defined as *four* levels, with segment #1 (the largest segment) using two of the levels. A four-level test element requires three columns in the design, so the A, B, and AB interaction columns were used to separate out the segmentation effect, as shown below.

**Three Customer Segments as 4-level Test Element**

<b>Combinations</b>	<b>A</b>	<b>B</b>	<b>AB</b>	<b>Segment</b>	<b>Available names</b>
1	-	-	+	Segment 2	11,586
2	+	-	-	Segment 1	13,508
3	-	+	-		
4	+	+	+	Segment 3	8,966

After creating the test design, one element dropped out. The team planned to test a search box at the top of the e-mail, but it was too difficult to execute for the test, so it was eliminated and column E left empty. The remaining 12 elements plus the 4-level segment variable are shown below.

<b><u>Test Elements</u></b>	<b><u>(-) Control</u></b>	<b><u>(+) New Idea</u></b>
A Segment	(1, 2)	(1, 3)
B Segment	(1, 2)	(1, 3)
C Link to online catalog	No	Yes
D Background color	White	Blue
E (empty)		
F Design of e-mail	Simple	Stronger brand image
G Partner promotions	None	Offers from two partner companies
H Navigation bar on side	Current	Additional buttons
J Special-offer starburst	No	"Special e-mail offer" starburst
K Discount offer	15% off	No discount
L Free gift	None	Free pen & pencil set
M Products pictured	Few	Many
N "Valued customer" copy	Current	Stronger
O Cross-sell copy	Current	New copy
P Subject line	"Exclusive e-mail offer..."	"Special offer for our customers..."

**A and B: Segment** – The marketing team had defined three key customer segments, based primarily on recency of purchase. Segment 1 were those who had made a purchase online or in a store within the last 3 months; segment 2 within the last 6 months; and segment 3 within 12 months.

**C: Link to online catalog** – The team felt that an obvious “Shop our catalog online” button towards the bottom of the e-mail would encourage customers to visit the website.

**D: Background color** – All e-mails were sent with dark text on a white background. The creative director thought that changing to a blue background might help the e-mail stand out.

**F: Design of e-mail** – E-mails used a basic font with a small company logo at the top. They wanted to test a stronger brand image and colors, with a larger logo, and more stylized font.

**G: Partner promotions** – With brand-name products, the marketing team believed that promoting a couple of specific brands could help convince customers to make a purchase. They decided to promote two specific brands in two bright boxes under “Offers from our partners” at the bottom of the e-mail.

**H: Navigation bar on side** – E-mails currently went out with a sidebar similar to the navigation bar on their website, but with a shorter list of links. They didn’t want to test an e-mail without any sidebar, so instead they decided to test the current navigation bar versus one with more choices.

**J: Special-offer starburst** – Since e-mails were sent to a “select group of customers,” they wanted to play up the exclusivity with an eye-catching red star at the upper right stating “Special e-mail offer.”

**K: Discount offer** – The Internet director had gone back-and-forth between offering a special e-mail discount or not. He thought the discount helped, but never really quantified whether it pulled in enough sales to justify the lower margin.

**L: Free gift** – They had never before offered a free gift with online orders, so they thought it was worth a try (and other companies were doing it). They selected an attractive, but low-cost, pen and pencil set. After considering the offer for orders over \$50, they decided no minimum was a better, bolder choice.

**M: Products pictured** – Every e-mail focused on a selection of products—with pictures and prices—but they never knew how many were best. Some thought that a simple offer with just a few products would get people to respond faster, but others thought that a larger selection would give more people something of interest, so they decided to test a few products versus many products.

**N: “Valued customer” copy** - Their standard e-mail copy stated, “As a valued [company] customer, we would like to offer you these Internet-only specials.” They tested this against copy with a stronger message, adding a second sentence about how only their best customers get these special offers.

**O: Cross-sell copy** – The second copy change was designed to sell more products. A sentence was added to encourage people to order a variety of office supplies at once to lower shipping costs.

**P: Subject line** – The Internet director had been testing different e-mail subject lines. Currently, “Exclusive e-mail offer from [company]” was the winner. Since he knew the subject line was important, he wanted to test another version, “Special offer for our best customers.”

## Test Design

The consultant developed a “fractional-factorial” test design<sup>3</sup> (shown below). The matrix shows sample size and response data (often called conversion rate) for each test recipe.

Since each customer segment was randomly assigned to certain test cells based on +/- levels in columns A and B, the test is not completed balanced. In addition, after names were assigned to test cells, the final purge/merge (where addresses are double-checked and invalid e-mail addresses are removed) dropped some names from the test. In the end, only 34,060 names were used. Each version of the e-mail was sent to as few as 641 or as many as 1,515 customers.

Recipe																# names	# orders	Response Rate
	Segment A	Segment B	Link to online catalog	Background color	(empty)	Design of e-mail	Partner promotions	Navigation bar on side	Special-offer starburst	Discount offer	Free gift	Products pictured	"Valued customer" copy	Cross-sell copy	Subject line			
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1515	21	1.39%
2	+	-	-	-	-	+	+	+	+	+	+	-	-	-	-	883	6	0.68%
3	-	+	-	-	-	+	+	+	-	-	-	+	+	+	-	883	9	1.02%
4	+	+	-	-	-	-	-	-	+	+	+	+	+	+	-	1169	8	0.68%
5	-	-	+	-	-	+	-	-	+	+	-	+	+	-	+	1515	9	0.59%
6	+	-	+	-	-	-	+	+	-	-	+	+	+	-	+	781	9	1.15%
7	-	+	+	-	-	-	+	+	+	+	-	-	-	+	+	883	3	0.34%
8	+	+	+	-	-	+	-	-	-	+	+	-	-	+	+	1180	14	1.19%
9	-	-	-	+	-	-	+	-	+	-	+	+	-	+	+	1515	17	1.12%
10	+	-	-	+	-	+	-	+	-	+	-	+	-	+	+	883	2	0.23%
11	-	+	-	+	-	+	-	+	+	-	+	-	+	-	+	800	14	1.75%
12	+	+	-	+	-	-	+	-	-	+	-	-	+	-	+	1180	0	0.00%
13	-	-	+	+	-	+	+	-	-	+	+	-	+	+	-	1515	9	0.59%
14	+	-	+	+	-	-	-	+	+	-	-	-	+	+	-	883	10	1.13%
15	-	+	+	+	-	-	-	+	-	+	+	+	-	-	-	883	11	1.25%
16	+	+	+	+	-	+	+	-	+	-	-	+	-	-	-	815	9	1.10%
17	-	-	-	-	+	-	-	+	-	+	+	-	+	+	+	1515	17	1.12%
18	+	-	-	-	+	+	+	-	+	-	-	-	+	+	+	883	7	0.79%
19	-	+	-	-	+	+	+	-	-	+	+	+	-	-	+	690	5	0.72%
20	+	+	-	-	+	-	-	+	+	-	-	+	-	-	+	1091	16	1.47%
21	-	-	+	-	+	+	-	+	+	-	+	+	-	+	-	1178	18	1.53%
22	+	-	+	-	+	-	+	-	-	+	-	+	-	+	-	883	3	0.34%
23	-	+	+	-	+	-	+	-	+	-	+	-	+	-	-	883	11	1.25%
24	+	+	+	-	+	+	-	+	-	+	-	-	+	-	-	1180	4	0.34%
25	-	-	-	+	+	-	+	+	+	+	-	+	+	-	-	1515	8	0.53%
26	+	-	-	+	+	+	-	-	-	-	+	+	+	-	-	641	7	1.09%
27	-	+	-	+	+	+	-	-	+	+	-	-	-	+	-	883	8	0.91%
28	+	+	-	+	+	-	+	+	-	-	+	-	-	+	-	1180	13	1.10%
29	-	-	+	+	+	+	+	+	-	-	-	-	-	-	+	1318	14	1.06%
30	+	-	+	+	+	+	-	-	+	+	+	-	-	-	+	883	9	1.02%
31	-	+	+	+	+	-	-	-	-	-	-	+	+	+	+	883	13	1.47%
32	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1171	8	0.68%

## Executing the Test

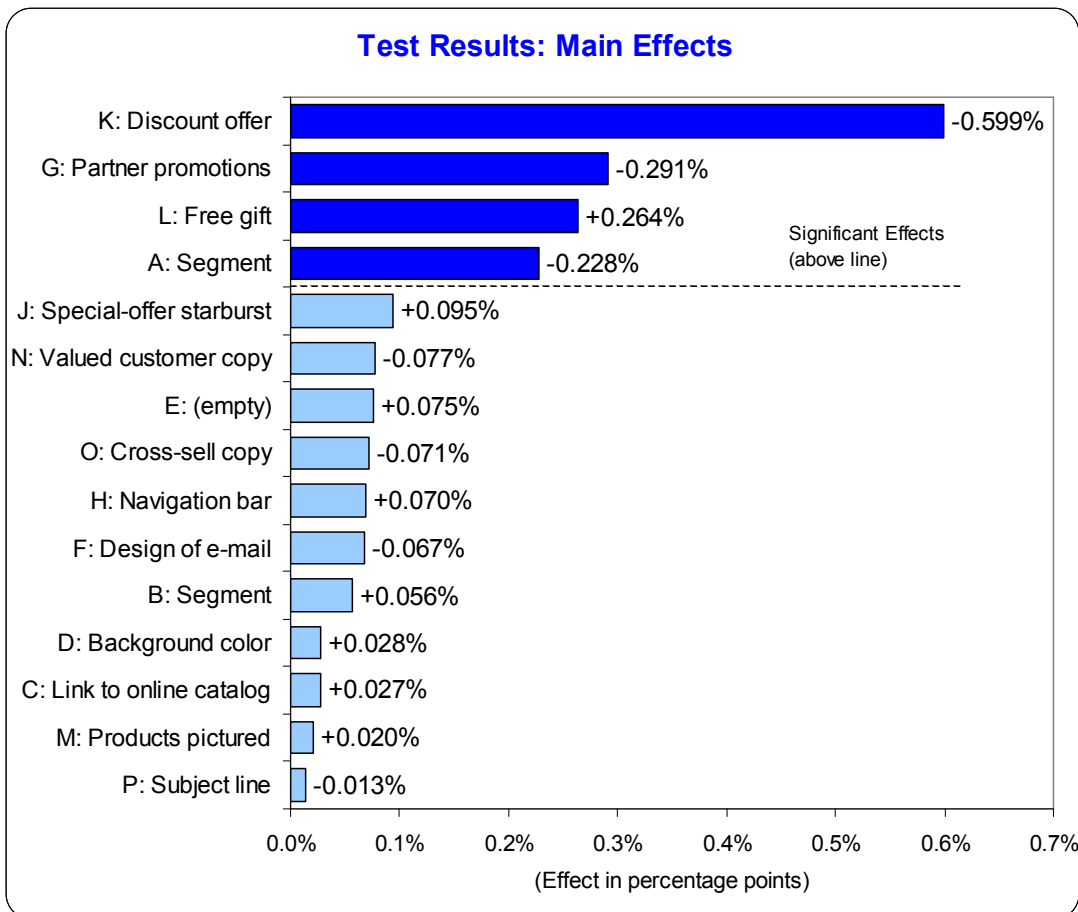
The creative team—made up of the creative director and one person who solely designed every e-mail—was a bit tentative about the test. The thought of creating 32 different e-mails for one drop was daunting. They also didn't know if all the required combinations would work from an artistic standpoint.

The consultant worked to minimize their concerns and workload. First he helped the team define clear, independent test elements that could work together in any combination. Then he provided “recipe sheets” listing the contents of each version. Following this, he met with the team to review every recipe, looking for challenging combinations and changing element definitions so all recipes could be created as simple cut-and-paste combinations. Finally, he worked with the creative team as they developed each version; he checked every creative proof to ensure compliance and consistency and solve any problems as they arose.

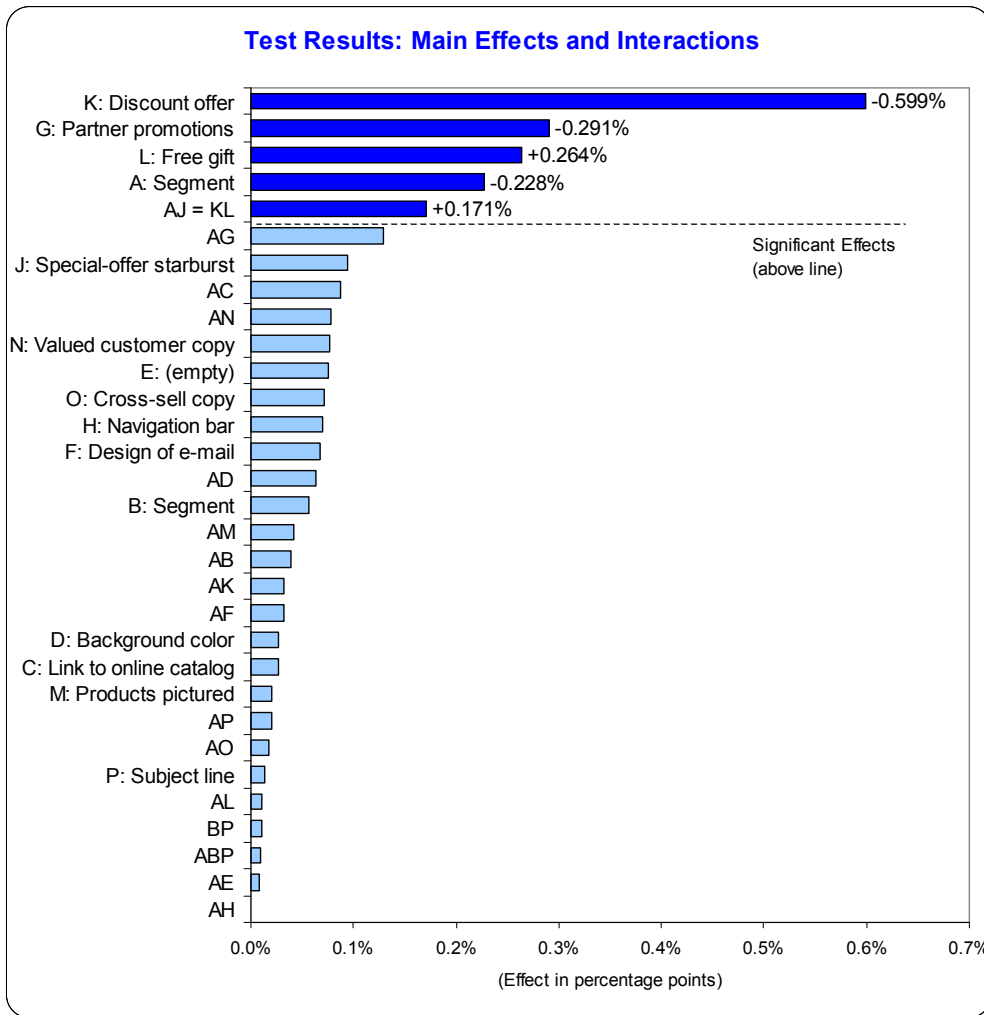
This work creating all 32 recipes added just two days to the marketing schedule. The team was surprised how smoothly things went once all test elements and recipes were clearly defined.

## Test Results

The test dropped on Tuesday and initial results were analyzed by Friday. Since the team wanted to increase the number of orders, the primary metric was response rate (also called conversion rate). Average order size was also analyzed to help assess profitability, but is not shown here. Main effects and two-way interactions are shown in the two bar charts, below.



Analyses of all interaction columns along with main effects show the effects, below.



### Significant Effects<sup>5</sup>

Average response rate for the test was 0.916%, with just 0-21 orders for each test cell and a total of only 312 orders. This was a small sample size in an unbalanced design with low response... and yet results were clear, robust, and insightful. Significant effects include:

#### K: Discount offer

The effect of -0.5993% means that eliminating the 15% discount results in a 52% drop in response. The team calculated that the loss of margin is more than covered by the increase in the number of orders.

#### G: Partner promotions

With a main effect of -0.2913%, the two partner offers in the e-mail *reduced* response by 25%, opposite what they expected. The team theorized that these additional offers may have confused the message and given customers too many disjointed offers to choose from.

#### L: Free gift

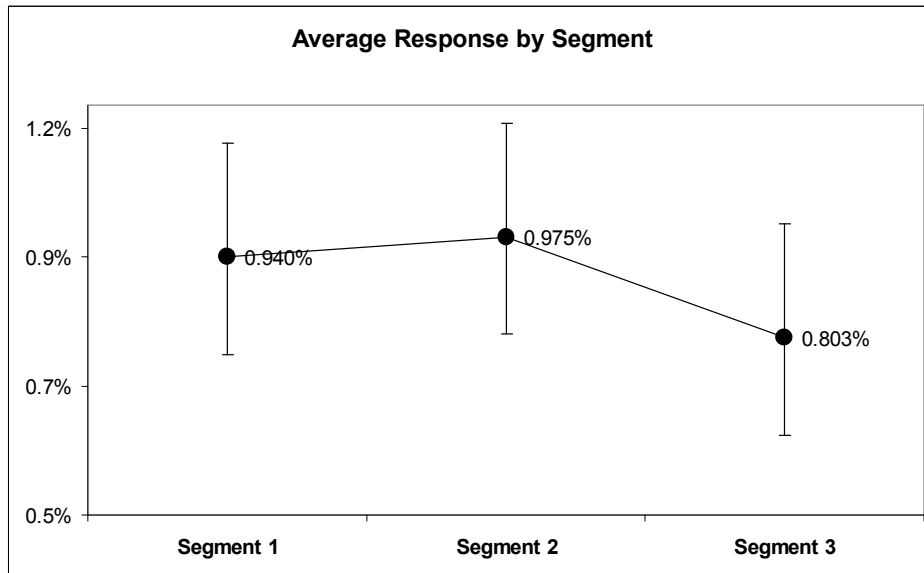
A positive effect of 0.2639% showed that the free pen and pencil set increased response nearly 23%. Analyzing profitability, the cost of the gift was easily covered by the increase in orders.

**A: Segment**

This test element was used as a block—the team knew that different segments responded differently, so they wanted to keep these differences from throwing off the test results. However, only one of the three segment columns was significant. Combining all three effects (A, B, and AB) to calculate average response by segment, gave the following results, summarized in the table and chart below:

Combinations	A	B	AB	Segment	Available names	Average Response
1	-	-	+	Segment 2	11,586	0.975%
2	+	-	-	Segment 1	13,508	0.789%
3	-	+	-			1.090%
4	+	+	+	Segment 3	8,966	0.803%

} 0.940%



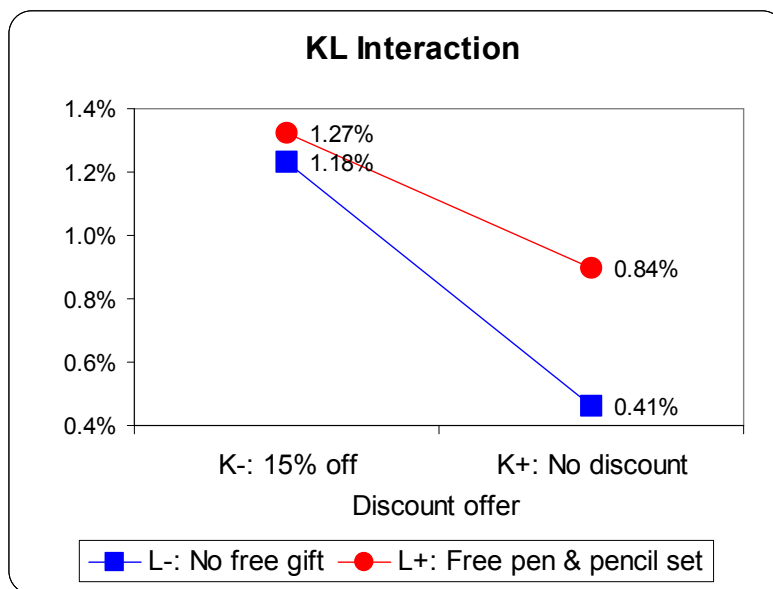
Analyzing all three levels together as a categorical variable shows that segment is not significant, even though element A on its own is. But data in the table, above, also raised a question: Why did half of segment 1 (A+B-) have a response rate lower than any other group, while the other half (A-B+) had the highest response rate of all?

After some investigation, the answer came down to a simple error in execution. The top half of segment 1 (i.e., the best customers) were separated out and then randomly placed in A-B+ test cells, while the bottom half were placed in A+B- test cells. This not only pointed out the risk of non-random assignment of names, but also showed that their segmentation model needs some refining—perhaps the best and worst recent buyers should be in different customer segments.

## KL Interaction

The final significant effect was the KL two-way interaction, with an effect of 0.1711%. Before explaining how this interaction affects results, it's worthwhile to take a step back and see where it came from<sup>4</sup>.

In this case, the KL interaction was by far the most likely. Both K and L are large significant effects, both are offer-related variables, and an interaction between the two can be logically explained (and pictured below).



This interaction supports the main effects: the 15% discount (K-, both points on the left) is always better and the free gift (L+, the top line) increases response over no free gift. However, the two-way interaction shows that both together—the 15% discount and the free gift—increase response less than the sum of both main effects.

The interaction can be understood by comparing both points on the left versus both points on the right. On the right (with no discount, K+), offering the free pen and pencil set gives a large jump in response versus offering no free gift—response more than doubles from 0.41% to 0.84%. In contrast, the points on the left show that, with the 15% discount (K-), the free gift increases response only slightly (from 1.18% to 1.27%).

Overall, this interaction shows that the 15% discount is great, the free gift is good, but both together are overkill—the free gift adds little to the benefit of the discount offer. These data helped them more accurately quantify their ROI on every combination of offers. Also, this gave the marketing team deeper insight into customer behavior, showing that one strong incentive is valuable, but additional incentives are probably unnecessary.

Offering the discount but no free gift, they predicted a response rate of 1.31%, or 1.41% with both offers together (versus response of 0.54% with neither offer). With these results, the Internet director decided to offer a discount more often, but sometimes switch to a free gift, depending on the e-mail campaign and the profitability of the customer segment.

## Conclusions

The Internet director was amazed by the depth and value of the results of this one test in one drop with just 34,060 names. He learned in one week more than he could with 8 months of testing using standard one-variable-at-a-time techniques.

With these results he decided to:

- Consistently offer the 15% discount (testing different discounts in future campaigns)
- Avoid the partner promotions that hurt response
- Use the special-offer starburst (J+): even though it was not significant, it was fairly close
- Offer a free gift every few e-mail campaigns to keep the offer fresh and sometimes offer it along with the discount to the highest-value customer segments (testing different gifts)
- Improve his segmentation model, adding more variables and splitting apart high-value and low-value recent buyers

Implementing these changes in the next campaign, response jumped to 1.54%, even higher than the prediction and more than 50% better than most past e-mail campaigns. The Internet director continued testing offers along with bolder creative changes, eventually achieving response rates consistently between 3% and 5% while adding more names in every drop.

After these results, the marketing team began testing changes in their catalog, retail stores, and regional advertising, continually squeezing greater profit from every marketing dollar. Every so often they found a major breakthrough, but more importantly, they frequently uncovered small changes that added up to a big bottom-line impact.

## Statistical Notes and Analyses

1. Size of the effect was calculated by rearranging the sample size equation into:

$$\text{Effect} = \sqrt{\frac{4 \cdot (Z_{\alpha/2} + Z_{\beta})^2 \cdot \bar{p} \cdot (1 - \bar{p})}{N}}$$

Where the overall sample size (across all test cells),  $N=35,000$ , the average response rate,  $\bar{p}=0.01$ ,  $Z_{\beta}=0$  (meaning the test has a 50-50 chance of seeing an effect of this size), and  $Z_{\alpha/2}=1.96$  (for 95% confidence). This gives an effect of 0.00208, or a 20.8% change.

2. The 32-run design would require greater effort for the marketing team to layout thirty-two different e-mails, but had important statistical advantages. Where a 16-run design would be resolution III (with two-way interactions fully confounded with main effects), the 32-run design is resolution IV: main effects are confounded with three-way interactions, but independent of two-way interactions. Since higher-order interactions are unlikely, this design reduces potential “confounding error” and also helps identify key two-way interactions.

3. The  $2^{(15-10)}$  fractional-factorial test design was based on a 32-recipe full-factorial design with elements F-P placed in ten of the 26 interaction columns, using design generators:

F = ABC, G = ABD, H = ABE, J = ACD, K = ACE, L = ADE, M = BCD, N = BCE, O = BDE, P = CDE

Ignoring four- and higher-order interactions results in a design (above) that can estimate the 15 main effects of elements A through P, 15 effects each containing 7 possible two-way interactions, and one effect containing only three-way interactions.

4. Initial analysis of the data—presented in the second Pareto chart above—shows 31 independent effects: 15 main effects, 15 two-way interactions, and one three-way interaction. Often, the default is to label the interactions with the first of all confounded interactions. For example, this significant interaction is labeled as AJ. However, this effect could actually be one or more of the seven confounded two-way interactions: AJ + BM + CD + EP + FG + KL + NO.

Initial analyses show that this column has a significant effect, but does not identify *which* interaction(s) is most likely. Here’s where marketing knowledge and statistical principles come together to help pinpoint the most likely interaction effects. Though interactions may be completely surprising, often they result from related test elements—variables located close together (like two elements on a direct mail envelope) or conceptually related (like price and offer variables).

5. These bar charts, also called Pareto charts, are the result of a series of statistical analyses using a variety of techniques. Objectives are to: (1) calculate all main effects, (2) determine how large an effect must be to be statistically significant, and (3) analyze interactions along with all significant main effects.

Main effects can be calculated simply by subtracting the average response rate of all 16 “-” recipes from the average of all 16 “+” recipes for each test element. This can be done by the marketing team with a pencil and paper, or more commonly, using Analysis of Variance (ANOVA).

Deciding what’s significant often progresses from simple to complex techniques:

(a) Look at the Pareto chart – Where is there a distinct “jump” in the effects? Is there a natural break that could signify the Line of Significance? The problem with this approach is that it’s just a guess. In reality, fewer or more effects, or nothing at all, may be significant.

(b) With a good measure of experimental error, ANOVA techniques will identify the significant effects if duplicate measurements for each recipe are available, or significance can be estimated using Lenth’s PSE, or by pooling small nonsignificant effects into the error term (but these last two add subjectivity to your estimates).

(c) Calculating standard error—basically rearranging the sample size equation to solve for the size of the effect—you can come up with a good measure of experimental error. However, this is not the most accurate approach.

(d) For response data, use logistic regression – especially with an unbalanced test, logistic regression (designed specifically for yes/no data) gives more accurate results, considering the number of orders and not just response rate (logically, a 1.00% response rate is probably more accurate with a sample size of 100,000 than 100). Also, regression is the best tool for the analysis of confounded interactions.

Test analysis is an iterative approach—looking at all effects, narrowing the list to significant main effects, and analyzing potential interactions and alternate models until the best model—using both statistical and marketing knowledge—is chosen. The final ANOVA and logistic regression models, including all components of the segmentation variable, are shown below (analyzed using MINITAB® Statistical Software):

**ANOVA - Factorial Fit: Response rate versus A, B, G, K, L, AB, KL**

Estimated Effects and Coefficients for Response Rate (coded units)

Term	Effect	Coef	SE Coef	T	P
Constant		0.9263	0.02823	32.81	0.000
A	-0.2275	-0.1138	0.02823	-4.03	0.000
B	0.0561	0.0280	0.02823	0.99	0.330
G	-0.2913	-0.1457	0.02823	-5.16	0.000
K	-0.5993	-0.2996	0.02823	-10.61	0.000
L	0.2639	0.1319	0.02823	4.67	0.000
A*B	-0.0397	-0.0199	0.02823	-0.70	0.488
K*L	0.1711	0.0855	0.02823	3.03	0.006

S = 0.159694 R-Sq = 88.68% R-Sq(adj) = 85.38%

Analysis of Variance for Response Rate (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	5	4.5485	4.5485	0.90969	35.67	0.000
2-Way Interactions	2	0.2468	0.2468	0.12339	4.84	0.017
Residual Error	24	0.6121	0.6121	0.02550		
Total	31	5.4073				

**Binary Logistic Regression: Orders, Names versus A, B, G, K, L, AB, KL**

Link Function: Logit  
Response Information

Variable	Value	Count
Orders	Success	312
	Failure	33748
Names	Total	34060

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-4.78157	0.0643679	-74.28	0.000			
A	-0.127779	0.0590138	-2.17	0.030	0.88	0.78	0.99
B	0.0238467	0.0590880	0.40	0.687	1.02	0.91	1.15
G	-0.163981	0.0577553	-2.84	0.005	0.85	0.76	0.95
K	-0.369910	0.0617801	-5.99	0.000	0.69	0.61	0.78
L	0.196689	0.0618118	3.18	0.001	1.22	1.08	1.37
AB	-0.0232412	0.0590546	-0.39	0.694	0.98	0.87	1.10
KL	0.157110	0.0618326	2.54	0.011	1.17	1.04	1.32

Log-Likelihood = -1744.333

Test that all slopes are zero: G = 60.822, DF = 7, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	12.0296	24	0.980
Deviance	15.3150	24	0.911
Hosmer-Lemeshow	2.0619	7	0.956
Brown:			
General Alternative	1.7748	2	0.412
Symmetric Alternative	1.7468	1	0.186