

DESIGNED EXPERIMENTS IN SERVICE QUALITY APPLICATIONS

Lou Johnson
Technical Training Specialist
Minitab, Inc.
1829 Pine Hall Road
State College, PA 16801
814-238-3280 x422
ljohnson@minitab.com

Gordon Bell
President
LucidView
80 Rolling Links Blvd.
Oak Ridge, TN 37830
888-582-4384 x707
gbell@lucidview.com

SUMMARY

“Almost any question can be answered, cheaply, quickly and finally, by a test campaign.”

– Claude Hopkins, *Scientific Advertising*, 1923

Experimental design and advertising testing were born of the same generation, but in worlds so far apart that it’s taken a lifetime for their divergent paths to come together. The basic concept behind in-market testing and design of experiments (DOE) is the same: a scientific way to prove which variables impact performance. Yet while advertising and marketing tests held firm to the one-variable-at-a-time approach, experimental design grew into a rich field with a wide array of test designs and advanced techniques offering the freedom to test more variables more efficiently.

Claude Hopkins became wealthy writing ads as an early leader in the field of direct-response marketing. Big city newspaper ads in the early 1900s often included a mail-in coupon. Newspapers with multiple runs offered a perfect channel for “split-run” testing. A different version of the ad would be placed in two different runs of the same day’s newspaper. Each ad version had a different “department number,” so when people mailed in the coupon, advertisers could simply count the number of coupons received from each ad to prove which was better. This offered a scientifically-valid way to test different headlines, pictures, prices, and any other idea for increasing response and sales.

Experimental design began in the farm fields of England. The techniques were applied during wartime and then in manufacturing operations. Sir Ronald Fisher published The Design of Experiments in 1935 (giving a name to this specialized field of statistics) and the techniques evolved over the years through academic research and industry application.

In this paper we will bring these two fields together to produce powerful experiments which have been designed to answer every transactional process owner’s quest to improve sales, revenue and customer satisfaction. The history of experimental design in service industries, the challenges of an ever changing market place, determining experimental run recipes, calculating an effective sample size and interpreting the results of the data analysis are reviewed.

INTRODUCTION:

THE CHASM BETWEEN STATISTICAL THEORY AND THE FRONT LINES OF THE MARKETPLACE

Since the early days of split-run newspaper ads, testing has remained a valuable part of most every marketing program. In direct mail, retail, Internet, call centers, sales, and customer service programs, market leaders are continually testing new ideas to increase sales and company performance. The use of DOE seems to ebb and flow as various quality programs come in and out of favor, but following the fundamentals of DOE remains the most efficient way to prove which variables drive performance.

As “testing” and “experimentation” grew in parallel, their paths seldom crossed. A smattering of technical papers in the 1970s and 1980s presented the theoretical potential for DOE in marketing and retail, but few, if any, experts were able to bridge the gap between concept and reality. Not until the mid-1990s—after seven decades—were these efficient experimental techniques applied in real-world marketing and service operations. In the authors’ experience, the biggest challenge over the years has been convincing the business world that something beyond one-variable-at-a-time testing is even possible (in apparent conflict with the “scientific method” taught in grade school). Even now, after “multivariable testing” (as DOE is often called) has entered the language of business, few market leaders have a good grasp of what that term means. At this point, surprisingly few market leaders are consistently using experimental design in marketing programs. This field remains fertile ground for greater education, experience, and application.

PROVEN SUCCESS IN SERVICE INDUSTRIES

Published successes have helped build interest in DOE/multivariable testing:

- A 15-factor supermarket test uncovered 6 changes that led to a 150% increase in product sales. (*Direct Marketing*, December 1997)
- A national carpet retailer increased their sales 20% by experimentally determining the sales person / customer interactions that optimized customer purchases. (*Lean Six Sigma for Service*, 2002)
- A 19-factor direct mail credit card test pinpointed 5 significant effects for a 15% jump in response rate. (*International Journal of Research in Marketing*, 2006)
- GE Capital saved over \$3MM by implementing the results of a 7 factor designed experiment studying methods to collect unpaid debt. (*Quality Engineering*, 2000)
- A global newspaper tested 11 creative and 4 price elements in one mail drop for a 41% increase in net response. Split-run tests would have shown few significant results and missed important interactions. (*presented by Financial Times at DMA07*)

Despite this success, Snee & Hoerl still declare in their book, *Six Sigma, Beyond the Factory Floor*, 2005, “It is our experience that DOE is the most under utilized Six Sigma tool in real economy applications.” There are several reasons design of experiments has still not reached its full potential in service quality applications.

CHALLENGES (AND SOLUTIONS) OF EXPERIMENTATION IN THE MARKETPLACE

The reason why experimental design remains under-used after seven decades is more than a simple lack of awareness. The transition from textbook statistics to real-world application has unique challenges far beyond those in manufacturing environments. Constant change in the marketplace, people as experimental units, and intangible factors all require unique attention, different strategies and, at times, different statistics.

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

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

The variability of factor effects on purchase behavior tends to increase as testing moves “upstream” from the point of purchase. Therefore, in-market testing can measure buyer behavior better than market research studies measuring opinion or intent. Even within in-market tests, variability increases as the test is executed farther back from the point of purchase. For example, in the retail environment, in-store tests tend to have less noise than advertising tests. A newspaper advertisement or circular reaches a broader audience but is also separated by a greater length of time and distance from the purchase decision.

Marketing tests can bring a nearly endless list of marketing-mix variables to be tested. In particular, creative changes—like the words, pictures, and layout of a direct mail package—can be defined and tested in numerous ways. Whereas a manufacturing process often has a discrete number of “knobs” that can be adjusted, the intangible nature of marketing programs removes many physical limits on the potential number of factors. For example, even a simple concept like “price” is very flexible. The absolute price (e.g., \$9.99) cannot be separated from the creative presentation: the number of times price is shown, along with the size and location, highlights and starbursts, and other ways of presenting that price point to the consumer. Because of the number of factors that may impact the service process response, a successful experimental design must not shy away from incorporating a large number of test variables and efficiently measuring their impact. Factorial designs, especially fractional factorial or screening designs were created with just that goal. This makes them particularly suited to success in a service / marketing environment.

FACTORIAL DESIGNS

Despite the hindrances mentioned above, the outcome of a successful experiment can be so clearly profitable that the benefits far outweigh the challenges. When senior executives see a large, measurable increase in sales as a direct result of DOE, they are easily sold on further testing. Given that an experimental design is easily torn apart under the strain of dynamic market forces, successful designs must be robust, efficient and produce clear, actionable results.

Factorial designs are easily created, analyzed and interpreted. For example, consider an experiment where the acceptance rate of a direct mail credit card offer is the response and the experimenter would like to determine the impact of three predictor variables: annual fee, interest rate on the outstanding balance, and the percent cash-back rate. A two-level full-factorial design requires selecting a high and low level for each of the three variables and testing all combinations of the high and low levels for each of the three variables at least one time (a total of $2 \times 2 \times 2$ combinations or 8 runs). One variable level combination (often called a “test recipe”) might be to offer no annual fee at a high interest rate of 16% and a low cash back rate of 0.5%. The responses from each of these eight combinations of predictor variable settings (recipes) can be used to estimate the main effect of each predictor and their interactions with each other. If experiment time was at a premium, a half fractional design or 4 of the eight combinations could be used to save experiment costs, but also limit your conclusions to just the main effects of each of the predictors. In some situations, this is an excellent tradeoff to make, in others, not.

Estimation of interaction effects is also a key benefit of using factorial designs. Interactions occur when the main effect of one variable is dependent on the setting of another variable. Price tests are a common scenario where interactions are important to estimate. For example, in a price test measuring grocery store magazine sales, the dampening effect of a higher asking price was offset by adding more copies of the magazine in the rack. The positive effect of the extra copies led to the same sales count as achieved by a lower asking price, therefore generating more revenue overall. When interactions are important, higher resolution factorial designs are selected because of their ability to clearly estimate the selected interactions.

Common factorial designs are orthogonal (zero correlation in predictor variable settings) and balanced (all variable levels equally tested). These include full-factorial, fractional factorial, and Plackett-Burman screening designs. Being orthogonal and balanced, they will produce the most valid results under less-than-ideal conditions. With these designs, all main effects are independent and select interactions can be analyzed (depending upon the chosen design and confounding scheme). In most marketing programs, each treatment combination can be costly to develop and execute. Therefore, one objective is to minimize the number of runs while still clearly estimating the desired effects. In this case, fractional factorial or screening designs are well suited to meet this objective.

Testing factors at two levels is much preferred to testing multiple levels of any factor. Most importantly, if the main goal is estimating the main effect and potential interactions of each predictor variable, testing more than two levels is essentially wasted effort. For example, if an annual fee is going to have a negative effect on the acceptance rate of a direct mail credit card offer, testing no fee and a moderate fee of \$35 / year will allow the estimation of that effect. Adding a \$20 fee level to the test would add many more run combinations and very little new results. Multi-level factors also add statistical complexity and can be difficult to interpret, especially if interactions exist. One must also consider that most marketing factors are discrete (bonus offer vs. no bonus offer, headline A vs. B), therefore lending themselves to an experiment designed and analyzed considering only two levels. A series of two-level experiments can be far more insightful than an attempt to test every possible idea in one larger multi-level test.

POWER

The power of an experimental data collection is the ability of the experiment to detect that a predictor variable has an effect on the response. The higher the power, the smaller an effect on the response the experiment will be able to detect. The power of the data collection is mainly a function of the number of samples taken. But if the response is a proportion, the power also has a lesser dependence on the average proportion anticipated in the response. For an effect to be statistically significant, the size of the effect must be larger than the calculated error in the effect estimate. When the response is a proportion (i.e. potential customers that actually make a purchase at the web store), the standard error in the effect is:

$$2 * \text{square root} (p (1-p) / N)$$

where p is the average proportion and N = samples size.

This simple formula can be used to estimate the number of samples needed to detect a given shift (or effect) in the overall response rate. For example, if the overall proportion of customers purchasing at the web store was 2% and the total test sample size was 16,000 customers entering the web store, this experiment would be able to detect a change in the purchase rate as small as +/- .15% from the overall rate of about 2% at the 95% confidence level. As 16,000 is a reasonable sample size in marketing tests and a 7.5% improvement in sales represents a significant improvement in revenue, this size experiment has good potential impact at a reasonable cost.

Another issue to consider is that market variability can be surprisingly high and unstable conditions add greater variability in experiments run over time. In-market experiments must have sufficient sample size and the right metrics to measure purchase behavior.

For example:

- A recent retail experiment included 18 factors tested in 500 stores. With 10+ stores in each test recipe, we could identify outliers and allow for some special causes to be removed during the test period.
- A 16-run direct mail test included 1 million people divided among all runs—an immense sample size, but necessary, considering the average response rate of only 0.25%.
- One 11-factor e-mail experiment was run over three weekly e-mail campaigns (with nearly 1 million names in each). Since every campaign included a different promotion, this allowed us to analyze differences among campaigns as well as significant results across all three.

CASE STUDY

To illustrate the implementation of a designed experiment, consider a major retailer who measured the effects of the characteristics of an email solicitation on the number of customers that visited the web store. The email characteristics included:

1. The subject line
2. The salutation such as; “Greetings!” or “Hi [customer name]”
3. Call to action; pitch to take the next step
4. A promotional give away such as entry in a \$100 drawing or a \$25 gift card
5. Closing statement

In all, 16 combinations of the five email factors were assembled in a factorial design that represented key combinations, but did not require testing all the combinations which would have wasted experimental time and effort with little added value. The 16 test recipes were randomly sent to potential customers of the website and the number that responded to each email by visiting the web store was recorded. The results were phenomenal. The response rate to the best solicitation was more than three times greater than that of the lowest level response. Among the most coveted group of customers, the response rate was 75% greater. If you had asked marketing experts to predict the email characteristics most likely to increase website traffic, most would have predicted the wrong answer. With a simple designed experiment, a higher rate of website traffic—in fact, an extremely higher rate than past history would have predicted—was achieved quickly and with minimal effort.

EFFICIENT TEST STRATEGY

Low cost, high-speed experiments are important. In dynamic markets, no one knows for sure how long important findings will apply. In addition, lead time and data lags can be quite large, so tests must be efficient, insightful, and actionable.

For example, direct mail programs often require three months to develop and produce a new mailing, plus three months before receiving about 90% of all responses. In addition, every version of the mailing requires additional creative time and a new printing, so costs can be high. In this case, testing many ideas at once is important while limiting the number of runs, so low-resolution fractional-factorial designs are frequently used (followed by a “refining” test in the next campaign to confirm results and perhaps do further testing).

Ultimately, the goal is to increase profitability as much and as quickly as possible. There’s no room for the “perfect” experiment if it takes too much time or money. Careful initial planning, an efficient design, sufficient sample size, and useful insights are what matter most.

CONCLUSION

Designed experiments have tremendous potential but are greatly under-utilized in the Service Quality industries. In particular, in-market testing – including retail, direct mail, Internet, and advertising testing – provides ample opportunity to leverage the experimental design and analysis techniques that have been developed in the past three decades. But the enormous potential comes with unique challenges. Far from textbook conditions, the “front lines” of marketing and service operations deal with the uncertainty of human behavior with textbook statistics often the first casualty. However, success is within reach. With the right statistical tools, good process knowledge, and a clear strategy, designed experiments can be completed, leading the way to improved sales and customer satisfaction.