Experiments and Quasi-experiments:

Methods for Evaluating Marketing Options

Hospitality managers could achieve greater success with marketing initiatives using experiments or quasi-experiments to test those initiatives.

BY ANN LYNN AND MICHAEL LYNN

Hospitality executives have available a number of different research methodologies and tools to aid them in decision making. Each methodology is valuable in its own way, but no single technique can provide all the answers to decision makers' questions. Exploratory research (such as focus groups and depth interviews[1]) and descriptive research (such as surveys[2] or naturalistic observations[3]) can provide insight and understanding about business problems and opportunities and thereby guide decision makers' search for promising courses of action. These techniques do not, however, allow researchers to draw conclusions about cause-and-effect relationships. Consequently, these techniques are of limited usefulness in revealing how effective a specific potential action will be. That is the province of causal research methods, such as choice modeling and experimentation, which can help decision makers draw conclusions about the effects, benefits, and influences of their prospective actions. Exploratory, descriptive, and causal research methods have a place in every functional area of hospitality businesses. In this article we focus on the use of causal-research methods in hospitality marketing.

Although systematic data on the use of different types of research in marketing are not available, several informal sources suggest that marketers often rely on exploratory and descrip-

[1] See: Robert J. Kwortnik, Jr., "Clarifying 'Fuzzy' Hospitality-management Problems with Depth Interviews and Qualitative Analysis," on pages 117–129 of this issue of *Cornell Quarterly*.

[2] See: Matthew Schall, "Best Practices in the Assessment of Hotel-guest Attitudes," on pages 51–65 of this issue of *Cornell Quarterly*.

[3] See: Kate Walsh, "Qualitative Research: Advancing the Science and Practice of Hospitality," on pages 66–74 of this issue of *Cornell Quarterly*.

tive research, but rarely use causal research. For example, a web-based search by topic of *Quirk's Marketing Research Review* from 1986 to 2001 indicated that this magazine has published 162 articles on focus groups and 59 articles on telephone interviewing and mail surveys, but only 22 articles on choice modeling (i.e., conjoint or trade-off analysis).[4] "Experiments" was not even listed as a search topic. Assuming that the frequency with which different research methods are written about in marketing-research magazines roughly reflects the frequency with which those methods are used by marketing researchers, the data from *Quirk's* would indicate that 90 percent of marketing research is exploratory or descriptive and only 10 percent is causal. Similar estimates were obtained from a query of the founders of two large firms engaged in marketing research for the hospitality industry. One estimated that 95 percent of hospitality-marketing-research expenditures are devoted to exploratory or descriptive research, while the other estimated that 80 percent of hospitality-research budgets are for exploratory or descriptive research.[5] It is clear to us that causal methods such as experiments are a rarity in hospitality-marketing research today.

We advocate the increased use of experiments and quasi-experiments in hospitality-marketing research.

In this article we advocate the increased use of experiments and quasi-experiments in hospitality-marketing research. The article is divided into three sections. Section one contains an explanation of why marketers should use causal research methods to evaluate the effects on consumers of different marketing actions. Section two contains a brief description of two causal-research methods—namely, true experiments and quasi-experiments—along with a discussion of their strengths and weaknesses. Section three contains a discussion of issues relating to conducting experiments and quasi-experiments and interpreting their results, which should give the reader an understanding of how to conduct and evaluate this type of research.

The Need for Causal Research

Ideally, hospitality marketers would first conduct exploratory and descriptive research to get an understanding of marketing problems or opportunities and would use this information to develop multiple courses of action that they believe will address those problems or capitalize on those opportunities. The proposed courses of action would then be systematically tested to discover whether they actually influence consumption behavior. Too often, however, marketers conduct only exploratory or descriptive research (as described above) and then develop just one course of action based on what they learn from those exercises. Kevin Clancy and Peter Krieg characterize this failure to develop and test several marketing options as a form of "death-wish marketing."[6] The problem with this practice is that the marketplace is so complex that no single course of action, even if well grounded in an understanding of the marketplace, is assured of producing the desired outcomes. In fact, marketers have a history of failing more than they succeed. Consider the following statistics compiled by Clancy and Krieg:

- the average brand loses market share each year,
- 90 percent of new products fail within three years,
- the average advertisement returns only 1 to 4 percent on the investment made in it,
- only 16 percent of trade promotions generate a profit, and
- the average firm satisfies less than 80 percent of its customers.[7]

[4] *Quirk's Marketing Research Review* can be searched at www.quirks.com/articles/search.asp.

[5] The experts queried were Stanley Plog, founder of Plog Research, and Peter Yesawich, president and CEO of Yesawich, Pepperdine & Brown. The differences in their estimates probably reflect differences in the work done by their respective firms.

[6] Kevin J. Clancy and Peter C. Krieg, *Counter-intuitive Marketing* (New York: Free Press, 2000).

[7] *Ibid.*

These statistics suggest that there is enormous room for improvement in market research. If marketers rigorously tested various marketing options before settling on a course of action, we believe they could identify where failure will occur before encountering it first hand.

On the rare occasions that marketers do test different marketing options or evaluate specific marketing actions already undertaken, they often use exploratory or descriptive research methods that are poorly suited to support conclusions about the proposed actions' effects on consumer behavior. Focus groups and surveys, for instance, are used to get consumers' opinions about advertisements, frequency programs, new product ideas, and other marketing options under consideration. Those options that consumers report liking best are then implemented or continued. One example of this approach can be found in the Harris Ad Research Service, which surveys a national sample of adults about how much they like various ads being run in the marketplace and how effective they think those ads are.[8] Companies who subscribe to this service are told how consumer attitudes and opinions about their ads compare to the average of consumer attitudes and opinions about the other ads being evaluated. Presumably, companies use this information by continuing ads that score well and by discontinuing ads that score poorly. Among the hospitality brands whose ads have been evaluated using this service in the past several years are Avis, Burger King, Domino's, Hertz, Holiday Inn, KFC, McDonald's, Pizza Hut, Priceline.com, and Red Lobster.[9]

One of the problems with this use of descriptive research to evaluate marketing options is that it is based on incorrect assumptions about consumer psychology. Using consumers' attitudes and beliefs to predict how they will react to certain marketing initiatives assumes that their attitudes and beliefs strongly affect their consumption behavior. However, psychologists have found that behavior is affected by many factors and that specific attitudes and beliefs are only weakly predictive of how people will behave in any given situation.[10] For example, attitudes towards an ad are only weakly related to purchase intentions and brand choice.[11] The weak link between atti-

Glossary of Terms

Quasi-experiments: A class of common field-research techniques in which at least one treatment is manipulated and there is at least one comparison. The difference between quasi- and true experiments is that in quasi-experiments consumers are not randomly assigned to treatments.

Random assignment: Assignment of consumers to treatments in such a way that each consumer has an equal chance of getting each treatment.

True laboratory experiment: A true experiment conducted in a model of the real world (a lab). Laboratory experiments are useful in basic research in consumer behavior because they can identify and explain the general conditions that influence consumer choices. While laboratory experiments are high in internal validity, they tend to be low in external validity.

True field experiment: An experiment conducted in the real world. Field experiments use random assignment, but do not attempt to control all factors extraneous to the ones being manipulated. Field experiments are useful for answering applied hospitality marketing questions because they have high internal validity and high external validity.

Type-1 error: Concluding that the treatments being tested had an effect when they really did not.

Type-2 error: Concluding that the treatments being tested had no effect when they really did.—*A.L. and M.L.*

[8] Sample Ad Track findings are reported in *USA Today* every Monday and can be found at www.usatoday.com/money/advertising/adtrack/index.htm.

[9] This claim is based on the list of Ad Track findings available at www.usatoday.com/money/advertising/adtrack/index.htm.

[10] See: David G. Myers, "Behavior and Attitudes," in *Social Psychology*, third edition (New York: McGraw-Hill, 1990), pp. 33–68; and A.W. Wicker, "Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects," *Journal of Social Issues*, Vol. 25 (1969), pp. 41–78.

[11] See: Stephen P. Brown and Douglas M. Stayman, "Antecedents and Consequences of Attitude toward the Ad: A Meta-analysis," *Journal of Consumer Research*, Vol. 19, June 1992, pp. 34–51; and Gabriel Biehal, Debra Stephens, and Eleonora Curlo, "Attitude toward the Ad and Brand Choice," *Journal of Advertising*, Vol. 21, September 1992, pp. 19–36.

tudes toward an advertisement and purchase behavior explains why popular ad campaigns, like Taco Bell's ads that feature a Chihuahua saying, "Yo quiero Taco Bell," often fail to increase sales.[12] Asking consumers to predict their own behavior is also unreliable. Psychologists have found that people are not aware of all the factors that affect their behavior and, most important, that they cannot accurately predict how they will react to events.[13] This is one reason that attempts to sell healthful or low-fat menu items in restaurants have generally failed despite consumers' reports that they want and will buy such items.[14]

To achieve company goals, marketers need to know what the effects of various marketing options or actions are on consumers' purchase behavior. While exploratory and descriptive research can provide information about consumer perceptions of, and attitudes toward, marketing options, these techniques cannot answer questions about how those options will affect consumer behavior. This is because the underlying causes of behavior are too complex to be accurately predicted from attitudes and opinions—even by consumers themselves. Fortunately, causal research methods can answer such questions. In the sections below we describe and discuss two causal research methods—namely, experiments and quasi-experiments.

Experiments, Quasi-experiments, and Their Limitations

Experiments are a type of research based on the following logic. If you identify two or more groups that are equivalent, expose those groups to different treatments, and subsequently observe differences between the groups on some dimension of interest, then you can reasonably conclude that those differences must be caused by the treatments. The following are characteristics of true experiments: (1) at least one treatment group and one comparison group, (2) at least one outcome measure, and (3) random assignment of subjects to treatments. Experiments can be used to test the effects of different prices, ad appeals, sales promotions, product changes, or any other marketing actions being considered on consumer attitudes and, most important, behavior.

For example, an experiment examining the effects of two menu designs on a restaurant's sales might randomly assign dining parties to receive one of the two menus by flipping a coin. A party sees one menu design when heads comes up and the other menu design when tails comes up. By keeping track of which dining parties received which menu design, the experimenter can compare the average check achieved with each menu design. Assuming that the samples involved are large, random assignment of dining parties distributes their characteristics evenly across the different groups and ensures that the groups seeing each menu design (or treatment) really are comparable. Thus, any subsequent difference in average check observed between the treatment groups can be attributed to the menu designs, and the experimenter can be confident that she knows which of the two designs will produce the largest sales in that restaurant.

True experiments with random assignment to treatments are sometimes impossible or impractical. In this case, one can conduct a quasi-experiment, a procedure that has the following characteristics: (1) at least one manipulated treatment group and one comparison group[15] (2) at least one outcome measure, and (3) nonrandom assignment of subjects to treatments. For example, a restaurant-chain executive may want to test the effects on sales of a proposed renovation of the chain's restaurants. Randomly assigning units within the chain to renovation and nonrenovation treatment groups would not be practical because renovating enough restaurants to make such random assignment meaningful would be too costly. In such a case, one could

[12] Christine R. McLaughlin, "Animals Gone Commercial: Do They Sell Products?," as viewed at animal.discovery.com/convergence/commercials/marketing_print.html.

[13] See: David G. Myers, "Intuition: The Power and Limits of Our Inner Knowing," in *Exploring Social Psychology* (New York: McGraw-Hill, 1994), pp. 23–30; and R.E. Nisbett and T.D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, Vol. 84 (1977), pp. 231–259.

[14] See: Amy Oplinger, "Survey Says ... ," *Praxis*, Fall 1998–Winter 1999, pp. 82–85; and Wilbert Jones, "New Wealth from Health," *Restaurant Hospitality*, Oct. 1999, pp. 98–102.

[15] The comparison group in both experiments and quasi-experiments can be a different group or the treatment group at a different point in time.

use a quasi-experimental design to test the redesign options.[16] For example, the executive could (1) identify a pair of units that are well matched on relevant characteristics such as customer demographics and sales, (2) renovate one unit in the pair, and (3) compare the sales achieved by each unit. To the extent that the matched pair is similar on the characteristics that are most likely to affect the outcome variables, this quasi-experimental design provides a reasonable basis for conclusions about the effects of the renovation without the costs of renovating many units.

Although there are no theoretical limits to the size and complexity of experiments and quasi-experiments, practical considerations such as cost and the availability of suitable subjects generally restrict such studies to only a few conditions. One prominent experimental-marketing researcher reports that the average experiment he does has about three levels per variable studied.[17] Direct-mail experiments often involve as many as 12 different treatments, but direct-mail experiments are typically less expensive than others, so this represents the high end of practicable experiment size. Thus, experiments and quasi-experiments are primarily useful in selecting from among a relatively narrow range of options.[18]

[16] Interested readers can find a discussion of many quasi-experiment designs in: William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (New York: Houghton Mifflin, 2002).

[17] Eric Marder, *The Laws of Choice* (New York: Free Press, 1997).

[18] If a marketer is interested in identifying the effects of more than 12 different treatments, experimentation may be less cost-effective than choice modeling. This is not the place for a detailed description of choice modeling. However, *Cornell Quarterly* readers can find an overview of one type of choice modeling known as discrete-choice analysis in: Rohit Verma, Gerhard Plaschka, and Jordan J. Louvirre, "Understanding Customer Choices: A Key to Successful Management of Hospitality Services," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 43, No. 6 (December 2002), pp. 15–24. One thing to keep in mind is that experiments provide stronger evidence of cause-and-effect relationships than do choice models and, therefore, are preferable to choice models as long as their costs are not prohibitive. However, if an experiment is not possible or would be too expensive, then choice modeling is another causal method worth considering.

Adjusting Sample Size: A Temptation to Avoid

The procedure for deciding on a sample size described in the accompanying article is the correct method. However, the required sample sizes indicated by this method are usually large, and marketers often want to avoid the costs of working with such large samples. In those cases, marketers may be tempted to run an experiment with smaller sample sizes than the number recommended by standard procedure, analyze the results, and then add additional subjects if a practically meaningful but not statistically significant effect is found.

We advise against this two-step procedure for two reasons. First, the small initial sample sizes may result in chance reductions of the observed effect such that what is in reality a practically meaningful effect appears not to be so. In that case, marketers will not add subjects, and the statistical power needed to avoid this Type-2 error will not be available. Second, the decision to run the experiment with additional subjects only when there is a sizeable but not significant effect in the initial, small sample biases the final test with the larger sample and increases the probability of a Type-1 error. If marketers can afford the additional subjects required by the second step of this procedure, they should use that larger sample size in the first place.—*A.L. and M.L.*

Issues in Designing Experiments and Interpreting Their Results

The important issues that arise when designing marketing experiments and interpreting their results involve three types of validity—those being statistical-inference validity, internal validity, and external validity.[19] These three types of validity refer to the causal conclusions derived from an experiment.[20] Such a conclusion has statistical-inference validity if the experimenter can rule out chance as an explanation for the absence or existence of differences between treatment groups. Such a conclusion has internal validity if

[19] Academic researchers are concerned about a fourth type of validity—known as construct validity. A conclusion has construct validity if the variables being manipulated and measured in an experiment are correctly identified and labeled. This is a concern in basic science where researchers want to make conclusions about general, abstract constructs based on specific, concrete manipulations and measures. However, in applied marketing research, the variables researchers want to make conclusions about are generally defined by their operationalizations, so construct validity is not a potential problem.

[20] For a thorough discussion of these types of validity, see: Shadish *et al., op. cit.*

the experimenter can rule out nonchance causes other than the intended treatments as a source of differences between treatment groups. Finally, such a conclusion has external validity if it can be generalized beyond the experimental sample and context. The requirements for each of these types of validity are briefly discussed below along with the implications for marketers who are designing an experiment or interpreting its results.

Statistical-inference Validity

A marketer has established statistical-inference validity if he or she has ensured that chance or randomness does not explain the observed differences among the treatment groups. Statistical-inference validity is threatened by random (or chance) variations in the outcome variable of an experiment. Such variation can lead to one of two fundamental errors when interpreting the experiment's results. First, chance can increase differences among treatment groups and lead experimenters to conclude that the treatments had an effect when they really did not. This is known as "Type-1 error." Second, chance can decrease differences among treatment groups and lead experimenters to conclude that the treatments had no effect when they really did. This is known as "Type-2 error." Marketers can reduce these two threats to statistical-inference validity by selecting appropriate acceptable alpha levels, obtaining sufficient sample sizes, and reducing within-treatment-group variability. Each of these methods of reducing statistical error is described below.

Appropriate acceptable alpha levels. The alpha level for a study is the probability of making a Type-1 error. The actual alpha is reported on the output of statistical-analysis programs, and is sometimes referred to as the "p value." Marketers decide what probability of making a Type-1 error is acceptable and conclude that observed differences between treatments reflect real (nonchance) effects only when appropriate statistical tests indicate that the probability of making a Type-1 error is tolerable. The conventionally accepted alpha level is $p \leq .05$, meaning that the experimenter is willing to take no more than a 5-percent chance of accepting an observed effect as real when it is not. The reason for accepting some nonzero probability of making a Type-1 error is that lowering this probability increases the probability of making a Type-2 error. Thus, marketers must weigh the relative consequences of making a Type-1 or a Type-2 error when deciding on the acceptable alpha level for a study.

Another thing to keep in mind is that the probability of making a Type-1 error increases with the number of comparisons being made between treatment groups. Assuming that an experiment's treatments have no real effect, the probability of making a Type-1 error could be 5 percent when making one comparison between two treatments, but it may be 50 percent when using the same alpha level to make 10 different comparisons among multiple treatments. Thus marketers making separate comparisons among multiple treatment groups may want to select a more-stringent acceptable alpha level than would those making a single comparison or else choose a statistical analysis that helps control the error rate.

Sufficient sample sizes. Large samples are less susceptible to the vagaries of chance than are small samples. Thus, marketers can reduce the probability of making a Type-2 error (using a given alpha level) by increasing the sample size. However, this does not mean that marketers always need to use large samples. If real treatment effects are large or alpha levels are high, then Type-2 errors could be rare even with small samples. Since large samples are expensive to obtain, marketers should make sure they are needed before using them.

To save money, marketers should determine the sample size needed to keep the probabilities of Type-1 and Type-2 errors at desired levels. These calculations can be done by hand or on one of several available computer programs.[21] To calculate sample size, marketers must specify (1) the desired probability of Type-2 errors, (2) the real size of treatment effects, (3) the acceptable alpha level, and (4) the within-treatment variability in the outcome measure. Since the size of the real treatment effect is generally unknown (that is why an experiment is being conducted in the first place), marketers should decide what the smallest practically meaningful effect would be and use that as the effect size.

[21] A free, online program that calculates the needed sample size is available at http://calculators.stat.ucla.edu/powercalc/

Variability within treatments. While researchers hope for variability in the outcome measure between treatment groups (that is, the effect of the treatment), variability within treatment groups increases the chances of a Type-2 error. Variability within treatments is a measure of the consistency of subjects' responses to the treatment. Ideally, there is low variability within treatments, which indicates that subjects responded to the treatments in the same way. If there is high variability, then observed differences between treatments may be due to chance and not the manipulated treatment. So, one way marketers can reduce the probability of making a Type-2 error is to reduce differences in the outcome variable among the subjects within each treatment group. This can be accomplished by increasing the similarity of the subjects to one another and by increasing the uniformity of the conditions under which data are collected. For example, the restaurant-menu experiment described previously would have less variability in check size if it used only evening dining parties comprising one male and one female as subjects than if it included lunch and evening dining parties of all compositions. Of course, increasing the similarity of subjects and the uniformity of conditions can compromise the generalizability of the results, so one must take this potential shortcoming into account. More will be said about generalizability in a subsequent section.

Internal Validity

Internal validity is the strength with which one can conclude that the manipulated treatment caused the observed changes in the outcome measure. High internal validity occurs when all alternative explanations for the observed treatment effect have been ruled out. Confounded treatments are the threat to internal validity. Confounding occurs when the treatment groups differ prior to the treatments or when the treatments differ in more ways than intended. For example, an experiment in which men get one treatment and women get another confounds the treatment with the sex of the subject. In this case, the researcher cannot tell whether any difference between the treatment groups in the outcome variable was caused by the treatments or by the subjects' sex. Similarly, an experiment in which the experimenter must interact with the subject after personally delivering the treatment may confound the treatment with other experimenter actions. Psychological research has found that experimenters who knew what treatment subjects received and who subsequently interacted with the subjects often unintentionally behaved differently to those in the various treatment groups.[22] Confounding of this kind means that researchers cannot tell whether any differences between the treatment groups in the outcome variable were caused by the treatments or by the experimenter's actions. Such post-treatment confounding can be eliminated by keeping experimenters blind to the subject's treatment group. Pre-treatment confounding can be eliminated through random assignment of subjects to treatments and (barring random assignment) can be reduced through the matching of samples and the use of other quasi-experimental designs. Each of these latter means of promoting internal validity is discussed below.

When designing experiments, keep in mind that large samples are less susceptible to the vagaries of chance than are small samples.

Random assignment. Assigning subjects to treatments so that each subject has an equal chance of getting in each treatment group provides the greatest assurance that treatment groups are similar prior to the implementation of the treatments. As long as sample sizes are large, this random assignment distributes the subjects' characteristics evenly across the different treatment groups.[23] The larger the sample being randomly assigned, the greater the similarity between the resulting groups, but samples of 20 to 30

[22] Robert Rosenthal, *Experimenter Effects in Behavioral Research* (New York: Appleton-Century-Crofts, 1966).

[23] More precisely, random assignment makes groups comparable on the expected, pre-treatment level of the outcome variable. Essentially, it distributes the pre-treatment propensity to respond on the outcome variable evenly across groups.

subjects per treatment group are often sufficient to consider the different treatment groups as equivalent.[24]

Random assignment of individual consumers to treatments is easy when experiments are conducted in a laboratory or are conducted via post or e-mail. In those cases, the experimenter has control over which subjects get which treatment. In addition, many magazines and television-cable companies now have the ability to deliver distinct content to various (essentially random) subsets of their customers. This allows marketers to expose different people to different ads even though they are reading the same magazine or watching the same television show. Those people can then be contacted and asked to provide information used to compare the effectiveness of the different ads.

In some cases, random assignment of individuals to different treatments is not possible. For example, a restaurateur could not randomly assign individual dining parties in a field experiment that compares the effects on sales of playing two different styles of music over the sound system. In such cases, however, it is possible to use different units of analysis and to conduct true experiments by randomly assigning those units to treatments. A restaurateur could, for example, randomly determine which of two different styles of music are played each day for two months and could then compare the average daily sales under each style of music. In this case, any differences between days in the number and type of customers or other characteristics will be evenly distributed across the two treatment groups and any subsequent difference between the treatment groups in average daily sales can be safely attributed to the different styles of music. In general, researchers can assign many different units (e.g., individual consumers, multi-person dining parties, days, units of a restaurant chain) to treatment groups, but should make sure that those units are what are described by the outcome measures.[25]

If random assignment of individuals or other units of analysis is not practical, marketers can use a quasi-experimental design. To do this the marketer must try to anticipate all the variables that might affect the outcome variable and find naturally occurring units matched on those variables. Unfortunately, it is nearly impossible to anticipate all the relevant variables and find units that are perfectly matched thereon. Even if matched pairs could be found, it is possible that factors outside the experimenter's control could change one of the units during the course of the study and thereby create a new confound. For example, a competitor of one of two matched restaurants in a quasi-experiment could suddenly close, boost the other restaurant's sales, and confound the experiment. The internal validity of this simple quasi-experimental design falls far short of that for a true experiment with random assignment. There are a variety of more-complex quasi-experimental designs that help address different threats to internal validity, and marketers interested in conducting a quasi-experiment should consult experts about the options available. However, the internal validity of quasi-experiments is never as great as that of true experiments, so whenever practical, random assignment is the preferred method of assigning subjects to treatments.

External Validity

External validity is the extent to which an experiment's results apply or generalize to the real

[24] Any sample size that produces statistically significant results in an experiment with random assignment is sufficient for random assignment to have worked. As long as subjects are randomly assigned, any pre-treatment differences in propensity to respond on the outcome variable can be due only to chance. Statistical significance means that the post-treatment differences on the outcome variable are too large to be attributed to chance, so the sample size was (by definition) large enough to rule out pre-existing chance differences between treatment groups. Samples of 20 to 30 subjects per treatment are common in academic psychological experiments. However, psychologists are more concerned about the existence of a treatment effect than about its exact size. Marketers interested in reliable estimates of treatment-effect sizes will need to use samples larger than 20 to 30 subjects per treatment. In addition, marketers that are studying insensitive or highly variable outcome measures may need to use large sample sizes.

[25] In other words, if the unit being randomly assigned is days or restaurants then the outcome measure should be a daily or restaurant average. There are statistical techniques that allow researchers to correctly analyze experiments where the units of random assignment and the units of outcome measurement are different, but those are new and sophisticated statistical techniques that are likely to be beyond the typical executive or manager's ability to implement. Thus, we advise randomly assigning and measuring the same units.

marketing environment of interest. External validity is threatened by differences between the real-world and experimental samples, treatments, measured behavior, or contexts. A common example of such a threat can be found in test marketing of new entrée items by, for instance, McDonald's (e.g., the McRib sandwich). Restaurants (not only McDonald's) often label such items as special offers that are available "for a limited time only." The problem with that approach stems from the fact that the availability of the items will not remain limited if they are judged a success and permanently added to the menu. In other words, the experimental conditions differ from those to which the marketer wants to generalize the experimental results. This difference is important because limited availability increases demand for products.[26] Test markets that describe items as special offers available for a limited time generally inflate the demand for those items and do not provide good estimates of the demand that item would generate as a permanent addition to the menu.

The way to ensure external validity is to make the features of the experiment similar to the features of the situation to which the experimental results will be generalized. Marketers should draw a sample that is representative of the actual consumers of the product or service, deliver the treatments to subjects in the same way and in the same context that they will be delivered in the marketplace, and measure the same outcome variable that managers want to affect in the marketplace. However, it is expensive and difficult (if not impossible) to make experiments similar in all respects to real-world situations of interest. Thus, marketers must often conduct experiments that differ in some ways from the situations to which they want to generalize the experimental results. For example, marketers often settle for nonrepresentative samples or measure attitudinal outcome variables when it is consumers' behavior that they ultimately want to affect. How much these differences affect the generalizability of the results depends on the specifics of the case. Some things to keep in mind when evaluating the generalizability of results across samples and measures are discussed below.

People of different ages, sexes, and ethnicities, as well as people from different regions of the country or world, differ in terms of tastes, value priorities, and other factors that may affect their responses to marketing communications and offers. As a result, it is dangerous to draw conclusions about one group of people based on data about a different group of people. However, generalizing findings across groups of people can be reasonable when there are only small differences between the groups or when the differences that exist are unlikely to affect responses to the treatment. For example, researchers have found only small demographic and psychographic differences between the users of different brands within consumer-product and -service categories.[27] This suggests that marketers can run experiments on their own customers and safely generalize the results to all users of the product category. In addition, researchers have found that differences between African-American and Caucasian consumers do not affect their responsiveness to point-of-purchase displays or price discounts.[28] This suggests that marketers can generalize findings about the effects of these tactics among one ethnic group to the other. The important thing to keep in mind is that differences between two groups of people do not necessarily make it inappropriate to generalize results from one group to the other. Only when those differences affect

Differences between two groups of people do not necessarily make it inappropriate to generalize results from one group to the other.

[26] See: Laura A. Branon and Amy E. McCabe, "Time-restricted Sales Appeals: The Importance of Offering Real Value," *Cornell Hotel and Restaurant Administration Quarterly*, Vol. 42, No. 4 (August–September 2001), pp. 47–52.

[27] Rachel Kennedy and Andrew Ehrenberg, "There Is No Brand Segmentation," *Marketing Research*, Spring 2001, pp. 4–7.

[28] Corliss L. Green, "Differential Responses to Retail Sales Promotion among African-American and Anglo-American Consumers," *Journal of Retailing*, Vol. 71 (1995), pp. 83–92.

responsiveness to the experimental treatments is generalizability called into question.

The difficulty and expense of measuring purchase behavior in naturalistic experiments leads many marketing researchers to use self-reported attitudes, beliefs, or purchase intentions as outcome variables in experiments instead of using actual purchase behavior. As mentioned earlier, this practice is flawed because attitudes, beliefs, and intentions are weak predictors of actual behavior.[29] Consequently, treatments may affect attitudes, beliefs, and intentions, but not affect purchase behavior. Since purchase behavior is what marketers are ultimately trying to influence, that behavior should be used as the outcome variable in marketing experiments whenever possible.

When actual marketplace behavior cannot be measured in an experiment, researchers should measure consumer choice in an artificial situation that is structurally similar to the choice situation in the marketplace. Eric Marder developed an artificial-choice task (called STEP) that has similar choice options, information about each option, and ease of choosing each option as those that consumers face in the marketplace. He found that consumers' STEP choices closely parallel their marketplace choices.[30] Thus, choice tasks like STEP provide a reasonable alternative to marketplace choices when measuring the effects of marketing experiments.

Ann Lynn, Ph.D. (top photo), is an assistant professor in the department of psychology at Ithaca College (alynn@ithaca.edu). **Michael Lynn,** Ph.D. (bottom photo), is an associate professor at the School of Hotel Administration at Cornell University (wml3@cornell.edu).

Conclusion

To be as effective as possible, marketers should develop and test several potential courses of action before embarking on any of them. The best way to develop a variety of courses of action is to conduct exploratory or descriptive research. The best way to evaluate those options is to conduct a causal-research study that compares consumers' behavior when faced with various options. Experiments and quasi-experiments are underused, but are nevertheless powerful research tools that allow hospitality marketers to draw strong causal conclusions about the effects of pricing, design, and other changes on the amount of money customers spend, or the number of visits they make to an establishment.

Experiments should be designed to have statistical-inference validity, internal validity, and external validity. The perfect marketing experiment would have (1) a large sample size that was drawn randomly from the population the marketer wanted to make conclusions about, (2) random assignment of subjects to treatments, (3) treatments that are delivered in the same way and in the same contexts they would be delivered in the marketplace, and (4) measures of consumer choice, or purchase behavior, in the marketplace. Such an experiment would allow a marketer to identify with complete confidence the best option under consideration.

Unfortunately, practical considerations often require compromises in experimental design. Such compromises are not a reason to dismiss the experimental results out of hand. Instead, marketers should assess the level of threat to statistical inference, internal or external validity posed by a compromise in experimental design, and adjust their confidence in the experimental results accordingly. Even an imperfect experiment can provide useful input into the selection of marketing options as long as the marketer is aware of what conclusions the experiment can support and what conclusions it cannot. We hope that this article will increase such awareness among hospitality marketers and will encourage them to make greater use of this research tool. ■

[29] See: Myers, *op. cit.*; and A.W. Wicker, "Attitudes versus Actions: The Relationship of Verbal and Overt Behavioral Responses to Attitude Objects," *Journal of Social Issues*, Vol. 25, pp. 41–78.

[30] STEP measurement involves giving subjects a booklet that describes all the major competitors in a product category and instructs subjects to distribute ten stickers among the competing options to reflect the likelihood that the subjects would buy the products as described. Each product description is on a separate page of the booklet. Product descriptions include a brand name, a picture, a price and a summary of product attributes and benefits (taken from real promotional materials on that product). The number of STEP stickers a person gives a product is related to that person's subsequent purchase behavior. Furthermore, the average shares of STEP stickers products receive correlate at .92 with the products' actual market shares. See: Marder, *op. cit.*