**ELEVEN Experimentation**

Experiments can be conducted in the field or in some kind of laboratory, that is, in an artificial situation constructed by the researcher. The essence of any experiment is the attempt to arrange conditions in such a way that one can infer causality from the outcomes observed. In practice, this means creating conditions or treatments that differ in one precise respect and then measuring some outcome of interest across the different conditions or treatments. The goal is to manipulate conditions such that differences in that outcome (how many people buy, how many people choose) can then be attributed unambiguously to the difference between the treatment conditions. In a word, the experiment is designed to determine whether the treatment difference *caused* the observed outcomes to differ. More properly, we should say that with a well-designed experiment, we can be *confident* that the treatment difference caused the outcomes to differ. (The role of probability in hypothesis testing will be discussed in [Chapter 14](https://jigsaw.vitalsource.com/books/9781483387246/epub/OEBPS/s9781483398228.i985.xhtml).)

Experimentation should be considered whenever you want to compare a small number of alternatives in order to select the best. Common examples would include: (1) selecting the best advertisement from among a pool of several, (2) selecting the optimal price point, (3) selecting the best from among several product designs (the latter case is often referred to as a “concept test”), and (4) selecting the best website design. To conduct an experiment in any of these cases, you would arrange for equivalent groups of customers to be exposed to the ads, prices, or designs being tested. The ideal way to do this would be by randomly assigning people to the various conditions. When random assignment is not possible, some kind of matching strategy can be employed. For instance, two sets of cities can provide the test sites, with the cities making up each set selected to be as similar as possible in terms of size of population, age and ethnicity of residents, per capita income, and so on. It has to be emphasized that an experiment is only as good as its degree of control; if the two groups being compared are not really equivalent, or if the treatments differ in several respects, some of them unintended (perhaps due to problems of execution or implementation), then it will no longer be possible to say whether the key treatment difference caused the difference in outcomes or whether one of those other miscellaneous differences was in fact the cause. *Internal validity* is the label given to this kind of issue—how confident can we be that the specified difference in treatments really did cause the observed difference in outcomes?

Because experiments are among the less familiar forms of market research, and because many of the details of implementing an experiment are carried out by specialists, it seems more useful to give extended examples rather than walk you through the procedural details, as has been done in other chapters. The examples address four typical applications for experimentation: selecting among advertisements, price points, product designs, or website designs. Note, however, that there is another entirely different approach to experimentation which I will refer to as conjoint analysis. Although conjoint is in fact an application of the experimental method, the differences that separate conjoint studies from the examples reviewed in this chapter are so extensive as to justify their treatment in a separate chapter.

**Example 1: Crafting Direct Marketing Appeals**

This is one type of experiment that virtually any business that uses direct mail appeals, however large or small the firm, can conduct. (The logic of this example applies equally well to e-mail marketing, banner ads, search key words, and any other form of direct marketing.) All you need is a supply of potential customers that numbers in the hundreds or more. First, recognize that any direct marketing appeal is made up of several components, for each of which you can imagine various alternatives: what you say on the outside of the envelope (or the subject line in the e-mail), what kind of headline opens the letter (or e-mail), details of the discount or other incentive, and so forth. The specifics vary by context; for promotional e-mail offers, you can vary the subject line, the extent to which images are used, which words are in large type, and so forth; for pop-up ads, you can vary the amount of rich media versus still images, size and layout, and the like. To keep things simple, let’s imagine that you are torn between using one of two headlines in your next direct marketing effort:

1. “For a limited time, you can steal this CCD chip.”
2. “Now get the CCD chip rated #1 in reliability.”

These represent, respectively, a low-price come-on versus a claim of superior performance. The remainder of each version of the letter will be identical. Let’s further assume that the purpose of the campaign is to promote an inventory clearance sale prior to a model changeover.

To conduct an experiment to determine which of these appeals is going to produce a greater customer response, you might do the following. First, select two samples of, say, 200 or more customers from the mailing lists you intend to use, using formulas similar to those discussed in [Chapter 13](https://jigsaw.vitalsource.com/books/9781483387246/epub/OEBPS/s9781483398228.i960.xhtml), the sampling chapter. A statistician can help you compute the exact sample size you need (larger samples allow you to detect even small differences in the relative effectiveness of the two appeals, but larger samples also cost more). Next, you would use a probability sampling technique to draw names for the two samples; for instance, selecting every tenth name from the mailing list you intend to use for the campaign, with the first name selected assigned to treatment 1, the second to treatment 2, the third to treatment 1, and so forth. Note how this procedure is more likely to produce equivalent groups than, say, assigning everyone whose last name begins with A through L to treatment 1 and everyone whose last name begins with M through Z to treatment 2. It’s easy to see how differences in the ethnic backgrounds of A to L versus M to Z patronyms might interfere with the comparison of treatments by introducing extraneous differences that have nothing to do with the effectiveness or lack thereof of the two headlines under study.

Next, create and print two alternative versions of the mailing you intend to send out. Make sure that everything about the two mailings is identical except for the different lead-in: same envelope, mailed the same day from the same post office, and so forth. Be sure to provide a code so you can determine the treatment group to which each responding customer had been assigned. This might be a different extension number if response is by telephone, a code number if response is by postcard, different URL if referring to a website, and so forth. Most important, be sure that staff who will process these replies understand that an experiment is under way and that these codes must be carefully tracked.

After some reasonable interval, tally the responses to the two versions. Perhaps 18 of 200 customers responded to the superior performance appeal, whereas only 5 of 200 customers responded to the low-price appeal. A statistical test can then determine whether this difference, given the sample size, is big enough to be trustworthy (see [Chapter 14](https://jigsaw.vitalsource.com/books/9781483387246/epub/OEBPS/s9781483398228.i985.xhtml)). Next, implement the best of the two treatments on a large scale for the campaign itself, secure in the knowledge that you are promoting your sale using the most effective headline *from among those considered*.

**Commentary on Direct Marketing Example**

The example just given represents a field experiment: Real customers, acting in the course of normal business and unaware that they were part of an experiment, had the opportunity to give or withhold a real response—to buy or not to buy, visit or not visit a website, and so forth. Note the role of statistical analysis in determining sample size and in assessing whether differences in response were large enough to be meaningful. Note finally the assumption that the world does not change between the time when the experiment was conducted and the time when the actual direct mail campaign is implemented. This assumption is necessary if we are to infer that the treatment that worked best in the experiment will also be the treatment that works best in the campaign. If, in the meantime, a key competitor has made some noteworthy announcement, then the world has changed and your experiment may or may not be predictive of the world today.

In our example, the experiment, assuming it was successfully conducted, that is, all extraneous differences were controlled for, establishes that the “Rated #1 in reliability” headline was more effective than the “Steal this chip” headline. Does the experiment then show that quality appeals are generally more effective than low-price appeals in this market? No, the experiment only establishes that *this* particular headline did better than this *other*particular headline. Only if you did several such experiments, using carefully structured sets of “low-price” and “quality” headlines, and getting similar results each time, might you tentatively infer that low-price appeals in general are less effective for customers in *this* product market. This one experiment alone cannot establish that generality. You should also recognize that the experiment in no way establishes that the “Rated #1 in reliability” headline is the *best possible* headline to use; it only shows that this headline is better than the one it was tested against. The point here is that experimentation, as a confirmatory technique, logically comes late in the decision process and should be preceded by an earlier, more generative stage in which possible direct mail appeals are identified and explored so that the appeals finally submitted to an experimental test are known to all be credible and viable. Otherwise, you may be expending a great deal of effort merely to identify the lesser of two evils without ever obtaining a really good headline.

The other advantage offered by many experiments, especially field experiments, is that in addition to answering the question “Which one is best?,” they also answer the question “How much will we achieve (with the best)?” In the direct mail example, the high-quality appeal was responded to by 18 out of 200, giving a projected response rate of 9 percent. This number, which will have a confidence interval around it, can be taken as a predictor of what the response rate in the market will be. If corporate planning has a hurdle rate of 12 percent for proposed direct mail campaigns, then the direct mail experiment has both selected the best headline and also indicated that it is may not be worth doing a campaign using even the best of the headlines under consideration, as it falls below the hurdle.

Much more elaborate field experiments than the direct mail example can be conducted with magazine and even television advertisements. All that is necessary is the delivery of different advertising treatments to equivalent groups and a means of measuring outcomes. Thus, split-cable and “single-source” data became available in the 1990s (for consumer packaged goods). In split cable, a cable TV system in a geographically isolated market has been wired so that half the viewers can receive one advertisement while a different advertisement is shown to the other half. Single-source data add to this a panel of several thousand consumers in that market. These people bring a special card when they go shopping for groceries. It is handed to the cashier so that the optical scanner at the checkout records under their name every product that they buy. Because you know which consumers received which version of the advertisement, you can determine empirically which version of the ad was more successful at stimulating purchase. See Lodish et al. (1995) for more on split-cable experiments.

One way to grasp the power of experimentation is to consider what alternative kinds of market research might have been conducted in this case. For instance, suppose you had done a few focus groups. Perhaps you had a larger agenda of understanding the buying process for CCD chips and decided to include a discussion of alternative advertising appeals with a focus on the two headlines being debated. Certainly, at some point in each focus group discussion, you could take a vote between the two headlines. However, from the earlier discussion in the qualitative sampling chapter, it should be apparent that a focus group is a decisively inferior approach to selecting the best appeal among two or three alternatives. The sample is too small to give any precision. The focus groups will almost certainly give some insight into the kinds of responses to each appeal that may exist, but that is not your concern at this point. That kind of focus group discussion might have been useful earlier if your goal was to generate a variety of possible appeals, but at this point, you simply want to learn which of two specified appeals is best.

You could, alternatively, have tried to examine the attractiveness of these appeals using some kind of survey. Presumably, in one section of the survey, you would list these two headlines and ask respondents to rate each one. Perhaps you would anchor the rating scale with phrases such as “high probability I would respond to this offer” and “zero probability I would respond.” The problem with this approach is different from that in the case of focus groups—after all, the survey may obtain a sample that is just as large and projectable as the sample used in the experiment. The difficulty here lies with interpreting customer ratings obtained via a survey as a prediction of whether the mass of customers would buy or not buy in response to an in-the-market implementation of these offers. The problem here is one of external validity: First, the headline is not given in the context of the total offer, as it occurs within an artificial context (completing a survey rather than going through one’s mail). Second, there is no reason to believe that respondents have any good insight into the factors that determine whether they respond to specific mail offers. (You say you never respond to junk mail? Huh, me neither! Funny, I wonder why there is so much of it out there . . .)

Remember, surveys are a tool for description. When you want prediction—which offer will work best—you seek out an experiment. If it is a field experiment, then the behavior of the sample in the experiment is virtually identical, except for time of occurrence, to the behavior you desire to predict among the mass of customers in the marketplace. Although prediction remains irreducibly fallible, the odds of predictive success are much higher in the case of a field experiment than if a survey, or worse, a focus group were to be used for purposes of predicting some specific subsequent behavior.

**Example 2: Selecting the Optimal Price**

Pricing is a topic that is virtually impossible to research in a customer visit or other interview. If asked, “How much would you be willing to pay for this?” you should expect the rational customer to lie and give a low-ball answer! Similarly, the absurdity of asking a customer, “Would you prefer to pay $5,000 or $6,000 for this system?” should be readily apparent, whether the context is an interview or a survey. Experimentation offers one solution to this dilemma; conjoint analysis offers another, as described subsequently.

The key to conducting a price experiment is to create different treatment conditions whose *only* difference is a difference in price. Marketers of consumer packaged goods are often able to conduct field experiments to achieve this goal. Thus, a new snack product might be introduced in three sets of two cities, and only in those cities. The three sets are selected to be as equivalent as possible, and the product is introduced at three different prices, say, $2.59, $2.89, and $3.19. All other aspects of the marketing effort (advertisements, coupons, sales calls to distributors) are held constant across the three conditions, and sales are then tracked over time. While you would, of course, expect more snack food to be sold at the lower $2.59 price, the issue is *how much more*. If your cost of goods is $1.99, so that you earn a gross profit of 60 cents per package at the $2.59 price, then the low-price $2.59 package must sell at twice the level of the high-price $3.19 package (where you earn $1.20 per package) in order to yield the same total amount of profit. If the experiment shows that the $2.59 package has sales volume only 50 percent higher than the $3.19 package, then you may be better off with the higher price. Note how in this example, the precision of estimate supplied by experimentation is part of its attraction.

Business-to-business and technology marketers often are not able to conduct a field experiment as just described. Their market may be national or global, or product introductions may be so closely followed by a trade press that regional isolation cannot be obtained. Moreover, because products may be very expensive and hence dependent on personal selling, it may not be possible to set up equivalent treatment conditions. (Who would believe that the 10 salespeople given the product to sell at $59,000 are going to behave in a manner essentially equivalent to the 10 other salespeople given it to sell at $69,000 and the 10 others given it to sell at $79,000?) Plus, product life cycles may be so compressed that an in-the-market test is simply not feasible. As a result, laboratory experiments, in which the situation is to some extent artificial, have to be constructed in order to run price experiments in the typical business-to-business or technology situation. Here is an example of how you might proceed.

First, write an experimental booklet (or, if you prefer, construct a web survey) in which each page (screen) gives a brief description of a competitive product. The booklet or website should describe all the products that might be considered as alternatives to your product, with one page in the booklet describing your own product. The descriptions should indicate key features, *including price*, in a neutral, factual way. The goal is to provide the kind of information that a real customer making a real purchase decision would gather and use.

Next, select a response measure. For instance, respondents might indicate their degree of buying interest for each alternative, or how they would allocate a fixed sum of money toward purchases among these products. Various measures can be used in this connection; the important thing is that the measure provide some analogue of a real buying decision. This is why you have to provide a good description of each product to make responses on the measure of buying interest as meaningful as possible. Note that an advantage of working with outside vendors on this kind of study is that they will have resolved these issues of what to measure long ago and will have a context and history for interpreting the results.

Now you create different versions of the booklet or web survey by varying the price. In one example, a manufacturer of handheld test meters wished to investigate possible prices of $89, $109, and $139, requiring three different versions of the booklet. Next, recruit a sample of potential customers to participate in the experiment. This sample must be some kind of probability sample drawn from the population of potential customers. Otherwise the responses are useless for determining the best price. Moreover, members of the sample must be randomly assigned to the treatment groups. If you use a list of mail addresses or e-mail addresses and have some other information about customers appearing on these lists, it also makes sense to see whether the types of individuals who respond to the three treatments remain equivalent. If one type of buyer has tended to drop out of one treatment condition, for whatever reason, confidence in the results is correspondingly reduced. Finally, administer the experimental booklet. Again, this could be done by mail, on the web, or in person at a central site(s).

In this price example, to analyze the results, you would examine differences in the projected market share for the product at each price (i.e., the percentage of respondents who indicate an interest or who allocate dollars to the product, relative to response to the competitive offerings). To understand the results, extrapolate from the projected market shares for the product at each price point to what unit volume would be at that level of market share. For example, the $89 price might yield a projected market share of 14 percent, corresponding to a unit volume of 76,000 handheld test meters. Next, construct an income statement for each price point. This will indicate the most profitable price. Thus, the $109 price may yield a market share of only 12 percent, but this smaller share, combined with the higher margin per meter, may yield a larger total profit.

What you are doing by means of this experiment is investigating the *price elasticity* of demand, that is, how changes in price affect demand for the product. Of course, demand will probably be lower at the $109 price than the $89 price; the question is, Exactly how much lower? You might find that essentially no one is interested in the product at the highest price tested. In other words, demand is very elastic, so that interest drops off rapidly as the price goes up a little. The experiment in that case has averted disaster. Or (as actually happened in the case of the handheld test meter example) you might find that projected market share was almost as great at the $139 price as at the $89 price, with much higher total profit (which is only to say that demand proved to be quite inelastic). In this case, the experiment would have saved you from leaving a great deal of money on the table through overly timid pricing.

**Commentary on Pricing Example**

Whereas a direct mail experiment can be conducted by almost any businessperson with a little help from a statistician, you can readily understand why, in a semi-laboratory experiment such as just described, you might want to retain an outside specialist. Finding and recruiting the sample and ensuring as high a return rate as possible are nontrivial skills. Selecting the best response measure takes some expertise as well. In fact, a firm with a long track record may be able to provide the additional service of comparing your test results with norms accumulated over years of experience.

Note that your level of confidence in extrapolating the results of a laboratory experiment will almost always be lower than in the case of a field experiment. In the direct mail example, the experiment provided an exact replica of the intended campaign except that it occurred at a different point in time with a subset of the market. A much larger gulf has to be crossed in the case of inferences from a laboratory experiment. You have to assume that (1) the people in the obtained sample do represent the population of potential customers, (2) their situation of receiving a booklet and perusing it does capture the essentials of what goes on during an actual purchase, and (3) the response given in the booklet does represent what they would actually do in the marketplace if confronted with these products at these prices. By contrast, in the direct mail case, the sample can easily be made representative, inasmuch as the initial sample is the obtained sample; the experimental stimulus is identical with the real ad to be used; and the experimental response is identical to the actual response: purchase. Nonetheless, when field experiments are not possible, laboratory experiments may still be far superior to relying on your gut feeling—particularly when your gut feeling does not agree with the gut feeling of a respected and influential colleague.

Finally, the difficulty and expense of conducting a laboratory-style price experiment has pushed firms toward the use of conjoint analysis when attempting to estimate demand at various price points. This application of conjoint analysis makes use of simulations, as discussed in [Chapter 12](https://jigsaw.vitalsource.com/books/9781483387246/epub/OEBPS/s9781483398228.i937.xhtml).

**Example 3: Concept Testing—Selecting a Product Design**

Suppose that you have two or three product concepts that have emerged from a lengthy development process. Each version emphasizes some kinds of functionality over others or delivers better performance in some applications than in others. Each version has its proponents or partisans among development staff, and each version can be characterized as responsive to customer input obtained through earlier qualitative and exploratory research. In such a situation, you would again have two related questions: first, which one is best, and second, what is the sales potential of that best alternative (a forecasting question). The second question is important because you might not want to introduce even the best of the three alternatives unless you were confident of achieving a certain sales level or a certain degree of penetration into a specific competitor’s market share.

Generally speaking, the same approach described under the pricing sample can be used to select among these product designs. If you can introduce an actual working product into selected marketplaces, as consumer goods manufacturers can, then this is a field experiment and would be described as a *market test*. If you must make do with a verbal description of a product, then this is a laboratory experiment and would be described as a *concept test*. Whereas in the pricing example, you would hold your product description constant and vary the price across conditions, in this product development example, you would vary your product description across three conditions, while you would *probably* hold price constant. Of course, if your intent was to charge a higher price for one of the product designs to reflect its presumed delivery of greater functionality, then the price would vary along with the description of functionality, and what is tested is several different combinations of function + price.

Note, however, that the experimental results can only address the differences between the complete product descriptions as presented; if these descriptions differ in more than one respect, the experiment in no way tells you *which* respect caused the outcomes observed. Thus, suppose that you find that the high-functionality, high-price product design yields exactly the same level of customer preference as the medium-functionality, medium-price design. At least two explanations, which unfortunately have very different managerial implications, would then be viable: (1) the extra functionality had no perceived value and the price difference was too small to have an effect on preference; or (2) the extra functionality did stimulate greater preference, which was exactly balanced by the preference-retarding effect of the higher price. You would kick yourself for not having included a high-functionality, medium-price alternative, which would have allowed you to disentangle these effects. But then, once you decide it would be desirable to vary price and functionality across multiple levels, you would almost certainly be better off pursuing a conjoint analysis rather than a concept test. In a concept test you can examine only 2–4 combinations; whereas in a conjoint analysis you can examine hundreds of permutations.

When an experiment is planned, the cleanest and most certain inferences can be drawn when the product designs being tested differ in exactly one respect, such as the presence or absence of a specific feature or functionality. If both product design and price are issues, then it may be best to run two successive experiments, one to select a design and a second to determine price, or to run a more complex experiment in which both product design and price are systematically varied—that is, an experiment with six conditions composed of three product designs each at two price levels. At this point, however, many practitioners would again be inclined to design a conjoint study instead of an experiment. In terms of procedure, the product design experiment can be conducted by e-mail invitation, as in the price example, or, at the extreme, and at greater expense, at a central site using actual working prototypes and examples of competitor products.

**Commentary on Product Design Example**

There is an important problem with concept testing of the sort just described if we examine the situation faced by most business-to-business (B2B) and technology firms. Market tests, if they can be conducted, are not subject to the following concern, but we agreed earlier that in many cases B2B and technology firms cannot conduct market tests of the sort routinely found among consumer packaged-goods firms. The problem with concept tests, in B2B and technology contexts, lies with the second of the two questions experiments can address (i.e., not Which is best? but How much will we achieve with the best?). We may assume that B2B concept tests are neither more nor less capable than consumer concept tests at differentiating the most preferred concept among the set tested. External validity issues are certainly present, but they are the same as when consumers in a test kitchen read descriptions of three different kinds of yogurt and indicate which is preferred. The problem comes when you attempt to generate a sales forecast from the measure of preference used in the concept test. That is, generalizing the *absolute* level of preference is a different matter than generalizing the *rank order* of preferences for some set of concepts.

Consumer packaged-goods firms have a ready solution to this problem. Over the several decades that modern concept testing procedures have been in place, so many tests have been performed that leading research firms have succeeded in compiling databases that link concept test results to subsequent marketplace results by a mathematical formula. The data have become sufficiently rich that the concept test vendors have been able to develop norms, on a product-by-product basis, for translating responses on the rating scale in the laboratory into market share and revenue figures in the market. The point is that the rating scale results in raw form cannot be extrapolated in any straightforward way into marketplace numbers. Thus, consumers’ response to each tested concept may be summed up in a rating scale anchored by “Definitely would buy”/“Definitely would not buy.” For the winning concept, 62 percent of consumers might have checked “Definitely would buy.” Does this mean that the product will achieve 62 percent trial when introduced? Nope. Only when that 62 percent number is arrayed against past concept tests involving yogurt are we able to determine that, historically, this is actually a below-average preference rating for yogurt concepts (it might have been quite good for potato chips) that will likely only translate into a real trial rate of 29 percent given the database findings.

There is no straightforward algorithm for translating rating scale responses for a never-before-tested category into a sales forecast. As most B2B firms are new to such market research arcana as concept tests, this means that the necessary norms probably do not exist, making concept test results less useful than for consumer goods manufacturers. (By definition, an innovative technology product cannot have accumulated many concept tests, so the point applies in spades to B2B technology firms.) Thus, B2B and technology firms are encouraged to explore the possible uses of concept tests but cautioned to use them mostly for differentiating better from worse concepts. Absent historical norms, one of the most attractive features of experiments, which involves projecting the absolute level of preference for purposes of constructing a market forecast, is simply not feasible in B2B technology markets.

**Example 4: A–B Tests for Website Design**

For this example, assume your firm has a website, and that this site has an e-commerce aspect—browsers can buy something by proceeding through the site to a shopping cart, or if in a B2B context, can advance down the purchase path, whether by registering for access, requesting literature, signing up for a webinar, and so on. Another name for purchase path is purchase funnel, with the metaphor suggesting that large numbers of prospects take an initial step toward purchase, that a smaller number take the next step, and so forth, with only a small fraction of initial prospects continuing all the way to purchase. There is a drop-off at each step, and the size of the drop-off can be thought of as a measure of the inefficiency of your marketing effort. The better the design of your website—the more effective your web-based marketing effort—the greater the fraction of prospects who continue on to the next step. Although 100 percent is not a feasible goal, in most real-world contexts, pass-through is so far below 100 percent as to leave plenty of room for improvement.

Given this setup, the business goal is to improve the efficiency of your web-based marketing, and one path to this goal is to optimize your website design, as measured by the proportion of prospects, at any step, who continue on to the next step, for some specified purchase funnel. The outcome of any experimental effort, then, is straightforward: an improvement, relative to baseline, in the fraction of prospects who continue to the next step.

The complexity comes in the nature of the experimental design itself. Recall that in the direct mail example, there was the simplest possible experiment: a headline or subject line offered to one group in version A, and to another group in version B. And that’s the heart of any A–B test: exposing people to two versions of something, with an opportunity to observe differences in response. But now consider how much more complex an e-commerce website is, relative to a simple direct mail pitch. There will be a home or landing page that will contain text of several kinds, probably one or more images, and a series of links. These links will be part of a purchase funnel, designed to move the casual browser on to the next step, and providing different routes for doing so. Each of these routes makes use of secondary and tertiary landing pages, which again have links designed to move the prospect forward toward purchase or inquiry. If e-commerce is a primary goal of the website, the entire link structure can be thought of as a sort of Venus flytrap, designed to attract the casual browser buzzing around the web ever deeper into the purchase funnel, until he’s committed.

Another key difference relative to the direct mail example is that your website is in operation already, and must remain in operation. This fact has a number of implications. First, any experimentation must not be too radical or disruptive. This is not a matter of experimenting offline with an e-mail to a few hundred customers from your vast customer base, where you could try something wild and crazy. This website *is* your marketing effort, and it has to be the best it can be, yesterday, today, and every day. Second, there are literally countless ways to vary some aspect of your website: the phrasing of this introductory text, the choice of that image, whether to foreground this link, and so forth. Third, it would be cumbersome to test different manipulations offline, and few would believe that browsing behavior in the lab under observation would generalize exactly or tightly enough to casual browsing in a natural setting. Fourth, your current website design is the fruit of a long history of attempts to design the website in the best possible way. This suggests that the most natural structure for an A–B test would be to let the current site design, exactly as it is, constitute the “A” treatment, which will function as a baseline or control group, against which some novel redesign can be compared. In this structure, there is in some sense only one treatment, the novel design element not now present on your website, which you want to test to see if it is an improvement.

Given all of the above, A–B testing in the context of website design consists of programming the host computer system to randomly serve up two versions of the website as each browser comes upon it. These can be two versions of the primary landing page, or two versions of any secondary or tertiary landing page. The programming is simple and assignment to the A or B version truly is random, fulfilling one of the most important conditions of experimental design. If the site has any reasonable amount of traffic, you’ll have an adequate sample size within a day or two. If the site has a lot of traffic, you can control the risk that the B version is actually worse by serving it up to only a fraction of browsers. Since you automatically have a good probability sample, once the B version has a big enough sample, the comparative statistical analysis can proceed, even if the B version was served up to only hundreds and the A version to tens of thousands (see [Chapter 13](https://jigsaw.vitalsource.com/books/9781483387246/epub/OEBPS/s9781483398228.i960.xhtml) on quantitative sampling).

If the B version is found to produce more of the desired outcome, such as a higher percentage of clickthrough to the next step in the purchase funnel, then the next day you tell your website programmers to retire the A version and serve up the B version from now on to all browsers. The new B version then becomes the baseline against which future attempts at optimization are tested.

**Commentary on the A–B Test Example**

Although an A–B test in principle is simply an experiment with two treatments, no different in kind from the direct mail example with which I began, its application to website design produces a noteworthy and distinctive twist on conventional experimentation. To some extent it shares elements with Big Data, most notably its scale, it application in real time, and even its potential automaticity. That is, in the example, the website redesign was portrayed as a one-off effort. Someone had a bright idea for improving the current site, and this new B version was approved for a test. But this limitation is not intrinsic. I can equally well imagine that the website has a crew of designers who are constantly thinking up new twists on the current design—as might be the case if the site in question is a major e-commerce site. In that event, one can imagine that there are four candidate images for the primary landing page, each one a really “good” image, so far as one member of the design crew is concerned. For that matter, the design philosophy may be that the key image on the landing page needs to change weekly or monthly, for variety’s sake. In that case, A–B, or A–B–C–D tests, may be running every week, as different images are continually vetted. It’s not that complex a programming task to set up the website hosting so that experimentation is ongoing—that is, so that B versions of one or another element of the site are constantly being tested on a fraction of browsers.

When experimentation is ongoing, so that large numbers of “experiments” are fielded every week on the website, a genuinely new kind of market research emerges. In the past, most design efforts in marketing were discrete efforts, occurring maybe once a year, or even less often. An example would be the design of a new television advertising campaign back in, say, 1989 (aka, “ancient times”). Even back then, multiple types of appeal, or two or three different ad executions, would typically be generated by the ad agency. The question of which one was best—the experimental question par excellence—was often decided politically, via the HIPPO rule (highest paid person’s opinion as to which one is best). If an experiment was done to decide the issue, it would be a laborious and often costly one-off effort, consuming months.

A vast gulf separates this old world from website design via A–B testing. Experiments in this domain are virtually cost-free, except for the salary cost for the designers who continually dream up new tweaks to the site design. Hundreds or thousands of experimental conditions can be run each year, and results can be obtained in a day or two. Opinion plays little role, except for deciding what tweaks are worth testing (assuming a surplus of ingenuity—else, everything gets tested). Experiments are field tests, and the results are unambiguous: the new tweak measurably improved site performance, or it did not. It’s a brave new world, in which experimentation becomes a mindset rather than a specialized skill.

**General Discussion**

Returning to the product design example, it may be instructive to examine once again the pluses and minuses of a controlled experiment as compared to conjoint analysis. The most important limitation of controlled experiments is that you are restricted to testing a very small number of alternatives. You also are limited to an overall thumbs-up or thumbs-down on the alternative as described. In conjoint studies, you can examine a large number of permutations and combinations. As we will see, conjoint analysis is also more analytic: It estimates the contribution of each product component to overall preference rather than simply rendering an overall judgment as to which concept scored best. Conversely, an advantage of controlled experiments is the somewhat greater realism of seeing product descriptions embedded in a context of descriptions of actual competitive products. Similarly, product descriptions can often be lengthier and more representative of the literature buyers will actually encounter, unlike conjoint analysis, where sparse descriptions are preferred in order to foreground the differences in level and type of product attribute that differentiate the design permutations under study.

**Future Directions**

Direct mail experiments such as the one described have been conducted since at least the 1920s. Experimentation is thus much older than conjoint analysis and on a par with survey research. Single-source data revolutionized the conduct of field experiments, but that was decades ago; the development of norms similarly revolutionized the conduct of concept tests, but that too occurred decades ago. The logic of conducting an experiment with e-mail or search engine key word purchases instead of “snail mail” is identical. Whether to do a concept test or a conjoint analysis remains a fraught question in many practical circumstances, but that’s been true for decades. Certainly, by some point in the 2000s, experimental stimuli became more likely to be administered by computer or over the web, but so what?

A reasonable conclusion with respect to experimentation, then, is that outside of website testing, nothing fundamental has changed. Worst of all, the knee-jerk reaction of most managers confronted with an information gap, certainly in B2B and technology markets, remains to do a survey or to interview customers. Hence, the power of experimental logic remains underappreciated. That stasis in mindset is far more important than the switch from paper to web presentation of experimental materials.

However, I do see one emerging trend that consists not so much of a new methodology as of a radical change in the economics and feasibility of experimentation, as described in the A–B testing example. The ability to deliver structured alternatives to a random sample of customers and observe any difference in outcome is the essence of experimentation. It follows that the cost, turnaround time, and feasibility of doing marketing experiments, in e-commerce, has plunged relative to what it costs to test different television ads, price points, or product concepts or any other marketing tool outside of an e-commerce context. This suggests that in the years to come, a mindset favoring experimentation may become a more crucial element in business (and in the career success of individual marketing professionals): In e-commerce, it’s cheaper, quicker, and easier to deploy this powerful research tool than ever before.

**Strengths and Weaknesses**

Experimentation has one crucial advantage that is so simple and straightforward that it can easily be overlooked or underplayed: Experiments promise causal knowledge. Experiments predict what will happen if you provide X functionality and/or price this functionality at Y. Although strictly speaking, even experiments do not offer the kind of proof available in mathematics, experiments provide perhaps the most compelling kind of evidence available from any kind of market research, with field experiments being particularly strong on this count. In short, experiments represent the most straightforward application of the scientific method to marketing decisions.

Experimentation has two subsidiary strengths. First, the structure of an experiment corresponds to one of the knottiest problems faced in business decision making: selecting the best from among several attractive alternatives. This is the kind of decision that, in the absence of experimental evidence, is particularly prone to politicking, to agonizing uncertainty, or to a despairing flip of the coin. In their place, experiments offer empirical evidence for distinguishing among good, better, and best. Second, experiments afford the opportunity to forecast sales, profit, and market share (again, this is most true of field experiments). The direct mail experiment described earlier provides a forecast or prediction of what the return rate, and hence the profitability, will be for the campaign itself. The pricing experiment similarly provides a prediction of what kind of market share and competitive penetration can be achieved at a specific price point, while the product design experiment provides the same kind of forecast for a particular configuration of functionality. These forecasts can be used to construct pro forma income statements for the advertising, pricing, or product decision. These in turn can be compared with corporate standards or expectations to make a go/no-go decision. This is an important advantage, inasmuch as even the best of the product designs tested may produce too little revenue or profit to be worthwhile. Other forecasting methods (e.g., extrapolation from historical data or expert opinion) are much more inferential and subject to greater uncertainty.

It must be emphasized that the predictive advantage of experiments is probably greatest in the case of field experiments. Laboratory experiments, when the goal is to predict the absolute level of response in the marketplace, and not just relative superiority, raise many issues of external validity. By contrast, in academic contexts where theory testing is of primary interest, laboratory experiments may be preferred because of considerations of internal validity. Academics routinely assume that nothing about the lab experiment (use of student subjects, verbose written explanations) interacts with the theory-based treatment difference, so that external validity can be assumed absent explicit evidence to the contrary. Practical-minded businesspeople can’t be so sure.

Experimentation is not without weaknesses. These mostly take the form of limits or boundary cases beyond which experimentation simply may not be feasible. For example, suppose there are only 89 “buyers” worldwide for your product category. In this case, you probably cannot construct two experimental groups large enough to provide statistically valid inferences and sufficiently isolated to be free of reactivity (reactivity occurs when buyers in separate groups discover that an experiment is going on and compare notes). In general, experiments work best when there is a large population from which to sample respondents. A second limit concerns products bought by committees or groups. It obviously does you little good if an experiment haphazardly samples fragments of these buying groups. You must either pick one kind of job role and confine the experiment to that kind, with consequent limits on your understanding, or find a way to get groups to participate in the experiment, which dramatically increases the costs and complexity.

More generally, experiments only decide between options that you input. Experiments do not generate fresh options, and they do not indicate the optimal possible alternative; they only select the best alternative from among those specified. This is not a problem when experiments are used correctly as the culmination of a program of research, but it can present difficulties if you rush prematurely into an experiment without adequate preparation. A related problem is that one can only select among a quite small number of alternatives. Conjoint analysis is a better route to go if you want to examine a large number of possibilities. Last, experiments can take a long time to conduct and can potentially tip off competitors, especially when conducted in the field.

**Dos and Don’ts**

* **Don’t** be overhasty in arranging to do an experiment. You really have to know quite a lot about customers and the market before an experiment can be valuable. If conducted prematurely, you risk getting wonderfully precise answers to the wrong question.
* **Don’t** let fear of costs prevent you from doing an experiment where appropriate. A laboratory experiment such as the price and product design examples described earlier may cost no more than a focus group study and considerably less than traditional survey research with a large national sample. E-mail tests, and concept tests using a website, can be very inexpensive.
* **Do** obsess about getting the details exactly right in your experimental design. The test groups have to be made as equivalent as possible, and the test stimuli have to differ in precisely those respects, and only those respects, under investigation.
* **Don’t** be afraid to be a pioneer. Experimentation is one of several areas of market research where business-to-business and technology firms tend to lag far behind consumer goods firms as far as best practice is concerned.
* **Don’t** expect brilliant new ideas or stunning insights to emerge from experiments. Experimentation does one narrow thing extremely well: It reduces uncertainty about whether a specific message, price change, or product design is superior to another and whether it surmounts a specified hurdle rate. Experimentation is confirmatory market research par excellence; it is not a discovery tool.

**Discussion Questions**

1. As discussed, there has been a great surge in A–B testing applied to website design. It is interesting to note that academic consumer research almost never uses a simple A–B test, except at the most early phases of pretesting for some larger experiment of interest. Instead, most experiments reported in the *Journal of Consumer Research* or the *Journal of Consumer Psychology* take the form of multi-way Analyses of Variance (ANOVA), in which two or more treatments, which may have more than two levels, are crossed.
	1. Why do A–B tests in a website context, when you could do A–B–C tests, or even A–B–C × X–Y tests? Are there really only two versions of the website that can be conceived at any one time?
	2. Might the focus on A–B tests have something to do with the difference between practitioner and academic concerns? Explain.
	3. A count in favor of practitioners is that A–B tests may be run repeatedly, even regularly, with successive tests of new “B” options. Does this change your answer to 1.a?
	4. Many websites get thousands of hits per day. This would allow for hundreds of subjects per cell in even a very complex experimental design. Say there are five versions of the central image, three ways of presenting site navigation links, and four options for the text lead-in on the home page. In a completely crossed design, that equals 60 experimental cells. Many websites could populate those cells with 500–1,000 browsers apiece, per day, more than enough for an acute statistical analysis. Explain why this more sophisticated multi-way ANOVA would or would not be superior to performing multiple A–B tests on successive pairs of cells.
2. Academic experiments in the marketing domain, especially laboratory experiments published in scholarly journals like *Journal of Consumer Research* or *Journal of Consumer Psychology*, are sometimes criticized for being unrealistic and artificial, or even irrelevant to managers. Critics point variously to the use of student subjects, paper and pencil measures of response in place of actual purchase behavior, environments that compel attention, and manipulations that could not be implemented outside of a university environment in which students are trained to read written instructions very, very carefully.
	1. Academics, naturally, defend their methods as appropriate for their purpose, which is to contribute new understanding at the theoretical level. A common defense is that “it’s not the data that generalize, but the theory.” This means that it doesn’t matter that the experimental data come from college sophomores filling out 7-point scales; the relationship between the theoretical concepts embodied in those scales is what will prove general, and provide new knowledge of what works in marketing. Evaluate this defense.
	2. A rebuttal would be that academic *psychologists* can be as theoretical as they wish, with no need to be relevant, but that someone claiming to do *marketing* science has to do experiments whose data will generalize to marketing contexts, that is, to the real circumstances with which marketing practitioners must deal. Otherwise, in what way is he or she a marketing scientist, as opposed to an academic psychologist? Discuss.
		1. *Extra Credit*: Salaries for marketing professors in a business school can be as much as twice as high as those for psychology professors—an extra hundred thousand dollars or more per year. Integrate this factor into your discussion of a and b above.
3. *Exercise*. Design an experiment for a business or in a product category you know well. Decide whether this will be an experimental test of product concepts, advertisements, or something else, and decide whether it will be a field or laboratory experiment. Your write-up needs to include: (1) statement of the decision problem, with background information as necessary; (2) specific research objectives; (3) research design, including the population from which you will sample, the sampling procedure, and the size of the initial and obtained samples (read the quantitative sampling chapter first); (4) a description of the two (or more) experimental treatments; (5) a description of the key outcome measure that will be used to assess the treatment effect(s); and (6) a description of any other questions that you think should be asked of experimental subjects as part of data collection.