SPOTLIGHT

ARTWORK Berndnaut Smilde, *Nimbus Green Room*, 2013
Digital C-type print, 75 x 102 cm/125 x 170 cm
Courtesy of the artist and Ronchini Gallery

The Discipline of Business Experimentation

Stefan Thomke is the William Barclay Harding Professor of Business Administration at Harvard Business School.

Jim Manzi is the founder and chairman of Applied Predictive Technologies, which provides software for designing and analyzing business experiments.

Increase your chances of success with innovation test-drives. *by Stefan Thomke and Jim Manzi*

Soon after Ron Johnson left Apple to become the CEO of J.C. Penney, in 2011, his team implemented a bold plan that eliminated coupons and clearance racks, filled stores with branded boutiques, and used technology to eliminate cashiers, cash registers, and checkout counters. Yet just 17 months after Johnson joined Penney, sales had plunged, losses had soared, and Johnson had lost his job. The retailer then did an about-face.

How could Penney have gone so wrong? Didn't it have tons of transaction data revealing customers' tastes and preferences?

Presumably it did, but the problem is that big data can provide clues only about the past behavior of customers—not about how they will react to bold changes. When it comes to innovation, then, most managers must operate in a world where they lack sufficient data to inform their decisions. Consequently, they often rely on their experience or

intuition. But ideas that are truly innovative—that is, those that can reshape industries—typically go against the grain of executive experience and conventional wisdom.

Managers can, however, discover whether a new product or business program will succeed by subjecting it to a rigorous test. Think of it this way: A pharmaceutical company would never introduce a drug without first conducting a round of experiments based on established scientific protocols. (In fact, the U.S. Food and Drug Administration requires extensive clinical trials.) Yet that's essentially what many companies do when they roll out new business models and other novel concepts. Had J.C. Penney done thorough experiments on its CEO's proposed changes, the company might have discovered that customers would probably reject them.

Why don't more companies conduct rigorous tests of their risky overhauls and expensive proposals? Because most organizations are reluctant to fund proper business experiments and have considerable difficulty executing them. Although the process of experimentation seems straightforward, it is surprisingly hard in practice, owing to myriad organizational and technical challenges. That is the overarching conclusion of our 40-plus years of collective experience conducting and studying business experiments at dozens of companies, including Bank of America, BMW, Hilton, Kraft, Petco, Staples, Subway, and Walmart.

Running a standard A/B test over a direct channel such as the internet—comparing, for instance, the response rate to version A of a web page with the response rate to version B—is a relatively uncomplicated exercise using math developed a century ago.

EXPERIMENT KOHL'S

The retailer set out to test the hypothesis that opening stores an hour later would not lead to a significant drop in sales.

But the vast majority (more than 90%) of consumer business is conducted through more-complex distribution systems, such as store networks, sales territories, bank branches, fast-food franchises, and so on. Business experimentation in such environments suffers from a variety of analytical complexities, the most important of which is that sample sizes are typically too small to yield valid results. Whereas a large online retailer can simply select 50,000 consumers in a random fashion and determine their reactions to an experimental offering, even the largest brick-and-mortar retailers can't randomly assign 50,000 stores to test a new promotion. For them, a realistic test group usually numbers in the dozens, not the thousands. Indeed, we have found that most tests of new consumer programs are too informal. They are not based on proven scientific and statistical methods, and so executives end up misinterpreting statistical noise as causation—and making bad decisions.

In an ideal experiment the tester separates an independent variable (the presumed cause) from a dependent variable (the observed effect) while holding all other potential causes constant, and then manipulates the former to study changes in the latter. The manipulation, followed by careful observation and analysis, yields insight into the relationships between cause and effect, which ideally can be applied to and tested in other settings.

To obtain that kind of knowledge—and ensure that business experimentation is worth the expense and effort—companies need to ask themselves several crucial questions: Does the experiment have a clear purpose? Have stakeholders made a commitment to abide by the results? Is the experiment doable? How can we ensure reliable results? Have we gotten the most value out of the experiment? (See the sidebar "Checklist for Running a Business Experiment.") Although those questions seem obvious, many companies begin conducting tests without fully addressing them.

Does the Experiment Have a Clear Purpose?

Companies should conduct experiments if they are the only practical way to answer specific questions about proposed management actions.

Consider Kohl's, the large retailer, which in 2013 was looking for ways to decrease its operating costs. One suggestion was to open stores an hour later on

Idea in Brief

THE PROBLEM
In the absence of sufficient data to inform decisions about proposed innovations, managers often rely on their experience, intuition, or conventional wisdom—none of which is necessarily relevant.

THE SOLUTION
A rigorous scientific test, in which companies separate an independent variable (the presumed cause) from a dependent variable (the observed effect) while holding all other potential causes constant, and then manipulate the former to study changes in the latter.

THE GUIDANCE
To make the most of their experiments, companies must ask: Does the experiment have a clear purpose? Have stakeholders made a commitment to abide by the results? Is the experiment doable? How can we ensure reliable results? Have we gotten the most value out of the experiment?

Monday through Saturday. Company executives were split on the matter. Some argued that reducing the stores' hours would result in a significant drop in sales; others claimed that the impact on sales would be minimal. The only way to settle the debate with any certainty was to conduct a rigorous experiment. A test involving 100 of the company's stores showed that the delayed opening would not result in any meaningful sales decline.

In determining whether an experiment is needed, managers must first figure out exactly what they want to learn. Only then can they decide if testing is the best approach and, if it is, the scope of the experiment. In the case of Kohl's, the hypothesis to be tested was straightforward: Opening stores an hour later to reduce operating costs will not lead to a significant drop in sales. All too often, though, companies lack the discipline to hone their hypotheses, leading to tests that are inefficient, unnecessarily costly, or, worse, ineffective in answering the question at hand. A weak hypothesis (such as "We can extend our brand upmarket") doesn't present a specific independent variable to test on a specific dependent variable, so it is difficult either to support or to reject. A good hypothesis helps delineate those variables.

In many situations executives need to go beyond the direct effects of an initiative and investigate its ancillary effects. For example, when Family Dollar wanted to determine whether to invest in refrigeration units so that it could sell eggs, milk, and other perishables, it discovered that a side effect—the increase in the sales of traditional dry goods to the additional customers drawn to the stores by the refrigerated items—would actually have a bigger impact on profits. Ancillary effects can also be negative. A few years ago, Wawa, the convenience store chain in the mid-Atlantic United States, wanted to introduce a flatbread breakfast item that had done well in spot tests. But the initiative was killed before the launch, when a rigorous experiment—complete with test and control groups followed by regression analyses—showed that the new product would likely cannibalize other more profitable items.

Have Stakeholders Made a Commitment to Abide by the Results?

Before conducting any test, stakeholders must agree how they'll proceed once the results are in. They should promise to weigh all the findings instead of cherry-picking data that supports a particular point of view. Perhaps most important, they must be willing to walk away from a project if it's not supported by the data.

When Kohl's was considering adding a new product category, furniture, many executives were tremendously enthusiastic, anticipating significant additional revenue. A test at 70 stores over six months, however, showed a net *decrease* in revenue. Products that now had less floor space (to make room for the furniture) experienced a drop in sales, and Kohl's was actually losing customers overall. Those negative results were a huge disappointment for those who had advocated for the initiative, but the program was nevertheless scrapped. The Kohl's example highlights the fact that experiments are often needed to perform objective assessments of initiatives backed by people with organizational clout.

Of course, there might be good reasons for rolling out an initiative even when the anticipated benefits are not supported by the data—for example, a program that experiments have shown will not substantially boost sales might still be necessary to build customer loyalty. But if the proposed initiative is a done deal, why go through the time and expense of conducting a test?

Checklist for Running a Business Experiment

Purpose

- Does the experiment focus on a specific management action under consideration?
- What do people hope to learn from the experiment?

Buy-In

- What specific changes would be made on the basis of the results?
- How will the organization ensure that the results aren't ignored?
- How does the experiment fit into the organization's overall learning agenda and strategic priorities?

Feasibility

- Does the experiment have a testable prediction?
- What is the required sample size? Note: The sample size will depend on the expected effect (for example, a 5% increase in sales).
- Can the organization feasibly conduct the experiment at the test locations for the required duration?

A process should be instituted to ensure that test results aren't ignored, even when they contradict the assumptions or intuition of top executives. At Publix Super Markets, a chain in the southeastern United States, virtually all large retail projects, especially those requiring considerable capital expenditures, must undergo formal experiments to receive a green light. Proposals go through a filtering process in which the first step is for finance to perform an analysis to determine if an experiment is worth conducting.

For projects that make the cut, analytics professionals develop test designs and submit them to a committee that includes the vice president of finance. The experiments approved by the committee are then conducted and overseen by an internal test group. Finance will approve significant expenditures only for proposed initiatives that have adhered to this process and whose experiment results are positive. "Projects get reviewed and approved much more quickly—and with less scrutiny—when they have our test results to back them," says Frank Maggio, the senior manager of business analysis at Publix.

When constructing and implementing such a filtering process, it is important to remember that experiments should be part of a learning agenda that supports a firm's organizational priorities. At Petco each test request must address how that particular experiment would contribute to the company's overall strategy to become more innovative. In the past the company performed about 100 tests a year, but that number has been trimmed to 75. Many test requests are denied because the company has done a similar test in the past; others are rejected because the changes under consideration are not radical enough to justify the expense of testing (for example, a price increase of a single item from $2.79 to $2.89). "We want to test things that will grow the business," says John Rhoades, the company's former director of retail analytics. "We want to try new concepts or new ideas."

Is the Experiment Doable?

Experiments must have testable predictions. But the "causal density" of the business environment—that is, the complexity of the variables and their interactions—can make it extremely difficult to determine cause-and-effect relationships. Learning from a business experiment is not necessarily as easy as isolating an independent variable, manipulating it, and observing changes in the dependent variable. Environments are constantly changing, the potential causes of business outcomes are often uncertain or unknown, and so linkages between them are frequently complex and poorly understood.

Consider a hypothetical retail chain that has 10,000 convenience stores, 8,000 of which are named QwikMart and 2,000 FastMart. The QwikMart stores have been averaging $1 million in annual sales and the FastMart stores $1.1 million. A senior executive asks a seemingly simple question: Would changing the name of the QwikMart stores to FastMart lead to an increase in revenue of $800 million? Obviously, numerous factors affect store sales, including the physical size of the store, the number of people who live within a certain radius and their average incomes, the number of hours the store is open per week, the experience of the store manager, the number of nearby competitors, and so on. But the executive is interested in just one variable: the stores' name (QwikMart versus FastMart).

The obvious solution is to conduct an experiment by changing the name of a handful of QwikMart stores (say, 10) to see what happens. But even determining the effect of the name change on those stores turns out to be tricky, because many other variables

Reliability

- What measures will be used to account for systemic bias, whether it's conscious or unconscious?
- Do the characteristics of the control group match those of the test group?
- Can the experiment be conducted in either "blind" or "double-blind" fashion?
- Have any remaining biases been eliminated through statistical analyses or other techniques?
- Would others conducting the same test obtain similar results?

Value

- Has the organization considered a targeted rollout—that is, one that takes into account a proposed initiative's effect on different customers, markets, and segments—to concentrate investments in areas where the potential payback is highest?
- Has the organization implemented only the components of an initiative with the highest return on investment?
- Does the organization have a better understanding of what variables are causing what effects?

may have changed at the same time. For example, the weather was very bad at four of the locations, a manager was replaced in one, a large residential building opened near another, and a competitor started an aggressive advertising promotion near yet another. Unless the company can isolate the effect of the name change from those and other variables, the executive won't know for sure whether the name change has helped (or hurt) business.

To deal with environments of high causal density, companies need to consider whether it's feasible to use a sample large enough to average out the effects of all variables except those being studied. Unfortunately, that type of experiment is not always doable. The cost of a test involving an adequate sample size might be prohibitive, or the change in operations could be too disruptive. In such instances, as we discuss later, executives can sometimes employ sophisticated analytical techniques, some involving big data, to increase the statistical validity of their results.

That said, it should be noted that managers often mistakenly assume that a larger sample will automatically lead to better data. Indeed, an experiment can involve a lot of observations, but if they are highly clustered, or correlated to one another, then the true sample size might actually be quite small. When a company uses a distributor instead of selling directly to customers, for example, that distribution point could easily lead to correlations among customer data.

The required sample size depends in large part on the magnitude of the expected effect. If a company expects the cause (for example, a change in store name) to have a large effect (a substantial increase in sales), the sample can be smaller. If the expected effect is small, the sample must be larger. This might seem counterintuitive, but think of it this way: The smaller the expected effect, the greater the number of observations that are required to detect it from the surrounding noise with the desired statistical confidence.

Selecting the right sample size does more than ensure that the results will be statistically valid; it can also enable a company to decrease testing costs and increase innovation. Readily available software programs can help companies choose the optimal sample size. (Full disclosure: Jim Manzi's firm, Applied Predictive Technologies, sells one, Test & Learn.)

How Can We Ensure Reliable Results?

In the previous section we described the basics for conducting an experiment. However, the truth is that companies typically have to make trade-offs between reliability, cost, time, and other practical considerations. Three methods can help reduce the trade-offs, thus increasing the reliability of the results.

Randomized field trials. The concept of randomization in medical research is simple: Take a large group of individuals with the same characteristics and affliction, and randomly divide them into two subgroups. Administer the treatment to just one subgroup and closely monitor everyone's health. If the treated (or test) group does statistically better than the untreated (or control) group, then the therapy is deemed to be effective. Similarly, randomized field trials can help companies determine whether specific changes will lead to improved performance.

The financial services company Capital One has long used rigorous experiments to test even the most seemingly trivial changes. Through randomized field trials, for instance, the company might test the color of the envelopes used for product offers by sending out two batches (one in the test color and the other in white) to determine any differences in response.

Randomization plays an important role: It helps prevent systemic bias, introduced consciously or unconsciously, from affecting an experiment, and it evenly spreads any remaining (and possibly unknown) potential causes of the outcome between the test and control groups. But randomized field tests are not without challenges. For the results to be valid, the field trials must be conducted in a statistically rigorous fashion.

Instead of identifying a population of test subjects with the same characteristics and then randomly dividing it into two groups, managers sometimes make the mistake of selecting a test group (say, a group of stores in a chain) and then assuming that everything else (the remainder of the stores) should be the control group. Or they select the test and control groups in ways that inadvertently introduce biases into the experiment. Petco used to select its 30 best stores to try out a new initiative (as a test group) and compare them with its 30 worst stores (as the control group). Initiatives tested in this way would often look very promising but fail when they were rolled out.

Now Petco considers a wide range of parameters—store size, customer demographics, the presence of nearby competitors, and so on—to match the characteristics of the control and test groups. (Publix does the same.) The results from those experiments have been much more reliable.

Blind tests. To minimize biases and increase reliability further, Petco and Publix have conducted "blind" tests, which help prevent the Hawthorne effect: the tendency of study participants to modify their behavior, consciously or subconsciously, when they are aware that they are part of an experiment. At Petco none of the test stores' staffers know when experiments are under way, and Publix conducts blind tests whenever it can. For simple tests involving price changes, Publix can use blind procedures because stores are continually rolling out new prices, so the tests are indistinguishable from normal operating practices.

But blind procedures are not always practical. For tests of new equipment or work practices, Publix typically informs the stores that have been selected for the test group. (Note: A higher experimental standard is the use of "double-blind" tests, in which neither the experimenters nor the test subjects are aware of which participants are in the test group and which are in the control. Double-blind tests are widely used in medical research but are not commonplace in business experimentation.)

Big data. In online and other direct-channel environments, the math required to conduct a rigorous randomized experiment is well known. But as we discussed earlier, the vast majority of consumer transactions occur in other channels, such as retail stores. In tests in such environments, sample sizes are often smaller than 100, violating typical assumptions of many standard statistical methods. To minimize the effects of this limitation, companies can utilize specialized algorithms in combination with multiple sets of big data (see the sidebar "How Big Data Can Help").

Consider a large retailer contemplating a store redesign that was going to cost a half-billion dollars to roll out to 1,300 locations. To test the idea, the retailer redesigned 20 stores and tracked the results. The finance team analyzed the data and concluded that the upgrade would increase sales by a meager 0.5%, resulting in a negative return on investment. The marketing team conducted a separate analysis and forecast that the redesign would lead to a healthy 5% sales increase.

As it turned out, the finance team had compared the test sites with other stores in the chain that were of similar size, demographic income, and other variables but were not necessarily in the same geographic market. It had also used data six months before and after the redesign. In contrast, the marketing team had compared stores within the same geographic region and had considered data 12 months before and after the redesign. To determine which results to trust, the company employed big

EXPERIMENT WAWA

A new flatbread did well in spot tests, but the chain killed it after rigorous experiments revealed it cannibalized other products.

How Big Data Can Help

To filter out statistical noise and identify cause-and-effect relationships, business experiments should ideally employ samples numbering in the thousands. But this can be prohibitively expensive or impossible. A new approach to merchandise assortment may have to be tested in just 25 stores, a sales-training program with 32 salespeople, and a proposed remodeling in 10 hotel properties. In such situations, big data and other sophisticated computing techniques, such as "machine learning," can help. Here's how:

Getting started

If a retailer wants to test a new store layout, it should collect detailed data (such as competitors' proximity, employees' tenures, and customer demography) about each unit of analysis (each store and its trade area, each salesperson and her accounts, and so on). This will become part of a big data set. Determining how many and which stores, customers, or employees should be part of the test and how long the test should run depends on the volatility in the data and the precision required for impact estimates.

Building a control group

In experiments involving small samples, correctly matching test subjects (such as individual stores or customers) to control subjects is essential and depends on the experimenter's ability to fully identify dozens or even hundreds of variables that characterize the test subjects. Big data feeds (complete transaction logs by customer, detailed weather data, social media streams, and so on) can assist in this. Once the characteristics are determined, a control group can be built that contains all elements of the test group except for what is being tested. This allows the retailer to determine whether the test results were influenced only by that one element—the new layout—or by other factors (demographic variances, better economic conditions, warmer weather).

Targeting the best opportunities

The same data feeds can be used to identify situations in which the tested program is effective. For example, the new store layout may work better in highly competitive urban areas but may be only moderately successful in other markets. By pinpointing these patterns, the experimenter can implement the program in situations where it works and avoid investments where the program may not generate the best ROI.

Tailoring the program

Additional large data feeds can be used to characterize program components that are more or less effective. For example, a retailer testing the effects of a new store layout can use data captured from in-store video streams to determine whether the new layout is encouraging customers to move through more of the store or is generating more traffic near high-margin products. Or the experimenter may find that moving items to the front of the store and putting in new shelves have a positive impact, but moving the sales registers disrupts checkouts and hurts profits.

data, including transaction-level data (store items, the times of day when the sale occurred, prices), store attributes, and data on the environments around the stores (competition, demographics, weather). In this way, the company selected stores for the control group that were a closer match with those in which the redesign was tested, which made the small sample size statistically valid. It then used objective, statistical methods to review both analyses. The results: The marketing team's findings were the more accurate of the two.

Even when a company can't follow a rigorous testing protocol, analysts can help identify and correct for certain biases, randomization failures, and other experimental imperfections. A common situation is when an organization's testing function is presented with nonrandomized natural experiments—the vice president of operations, for example, might want to know if the company's new employee training program, which was introduced in about 10% of the company's markets, is more effective than the old one. As it turns out, in such situations the same algorithms and big data sets that can be used to address the problem of small or correlated samples can also be deployed to tease out valuable insights and minimize uncertainty in the results. The analysis can then help experimenters design a true randomized field trial to confirm and refine the results, especially when they are somewhat counterintuitive or are needed to inform a decision with large economic stakes.

For any experiment, the gold standard is repeatability; that is, others conducting the same test should obtain similar results. Repeating an expensive test is usually impractical, but companies can verify results in other ways. Petco sometimes deploys a staged rollout for large initiatives to confirm the results before proceeding with a companywide implementation. And Publix has a process for tracking the results of a rollout and comparing them with the predicted benefit.

Have We Gotten the Most Value out of the Experiment?

Many companies go through the expense of conducting experiments but then fail to make the most of them. To avoid that mistake, executives should take into account a proposed initiative's effect on various customers, markets, and segments and concentrate investments in areas where the potential paybacks are highest. The correct question is usually not, What works? but, What works where?

Petco frequently rolls out a program only in stores that are most similar to the test stores that had the best results. By doing so, Petco not only saves on implementation costs but also avoids involving stores where the new program might not deliver benefits or might even have negative consequences. Thanks to such targeted rollouts, Petco has consistently been able to double the predicted benefits of new initiatives.

Another useful tactic is "value engineering." Most programs have some components that create benefits in excess of costs and others that do not. The trick, then, is to implement just the components with an attractive return on investment (ROI). As a simple example, let's say that a retailer's tests of a 20%-off promotion show a 5% lift in sales. What portion of that increase was due to the offer itself and what resulted from the accompanying advertising and training of store staff, both of which directed customers to those particular sales products? In such cases, companies can conduct experiments to investigate various combinations of components (for instance, the promotional offer with advertising but without additional staff training). An analysis of the results can disentangle the effects, allowing executives to drop the components (say, the additional staff training) that have a low or negative ROI.

Moreover, a careful analysis of data generated by experiments can enable companies to better understand their operations and test their assumptions of which variables cause which effects. With big data, the emphasis is on correlation—discovering, for instance, that sales of certain products tend to coincide with sales of others. But business experimentation can allow companies to look beyond correlation and investigate causality—uncovering, for instance, the factors causing the increase (or decrease) of purchases. Such fundamental knowledge of causality can be crucial. Without it, executives have only a fragmentary understanding of their businesses, and the decisions they make can easily backfire.

When Cracker Barrel Old Country Store, the Southern-themed restaurant chain, conducted an experiment to determine whether it should switch from incandescent to LED lights at its restaurants, executives were astonished to learn that customer traffic actually *decreased* in the locations that installed LED lights. The lighting initiative could have stopped there, but the company dug deeper to understand the underlying causes. As it turned out, the new lighting made the front porches of the restaurants look dimmer, and many customers mistakenly thought that the restaurants were closed. This was puzzling—the LEDs should have made the porches brighter. Upon further investigation, executives learned that the store managers hadn't previously been following the company's lighting standards; they had been making their own adjustments, often adding extra lighting on the front porches. And so the luminosity dropped when the stores adhered to the new LED policy. The point here is that correlation alone would have left the company with the wrong impression—that LEDs are bad for business. It took experimentation to uncover the actual causal relationship.

Indeed, without fully understanding causality, companies leave themselves open to making big mistakes. Remember the experiment Kohl's did to investigate the effects of delaying the opening of its stores? During that testing, the company suffered an initial drop in sales. At that point, executives could have pulled the plug on the initiative. But an analysis showed that the number of customer transactions had remained the same; the issue was a drop in units per transaction. Eventually, the units per transaction recovered and total sales returned to previous levels. Kohl's couldn't fully explain the initial decrease, but executives resisted the temptation to blame the reduced operating hours. They didn't rush to equate correlation with causation.

What's important here is that many companies are discovering that conducting an experiment is just the beginning. Value comes from analyzing and then exploiting the data. In the past, Publix spent

BEST PRACTICE PETCO

The specialty retailer ensures reliable results from its experiments by matching the characteristics of the control and test groups.

80% of its testing time gathering data and 20% analyzing it. The company's current goal is to reverse that ratio.

Challenging Conventional Wisdom

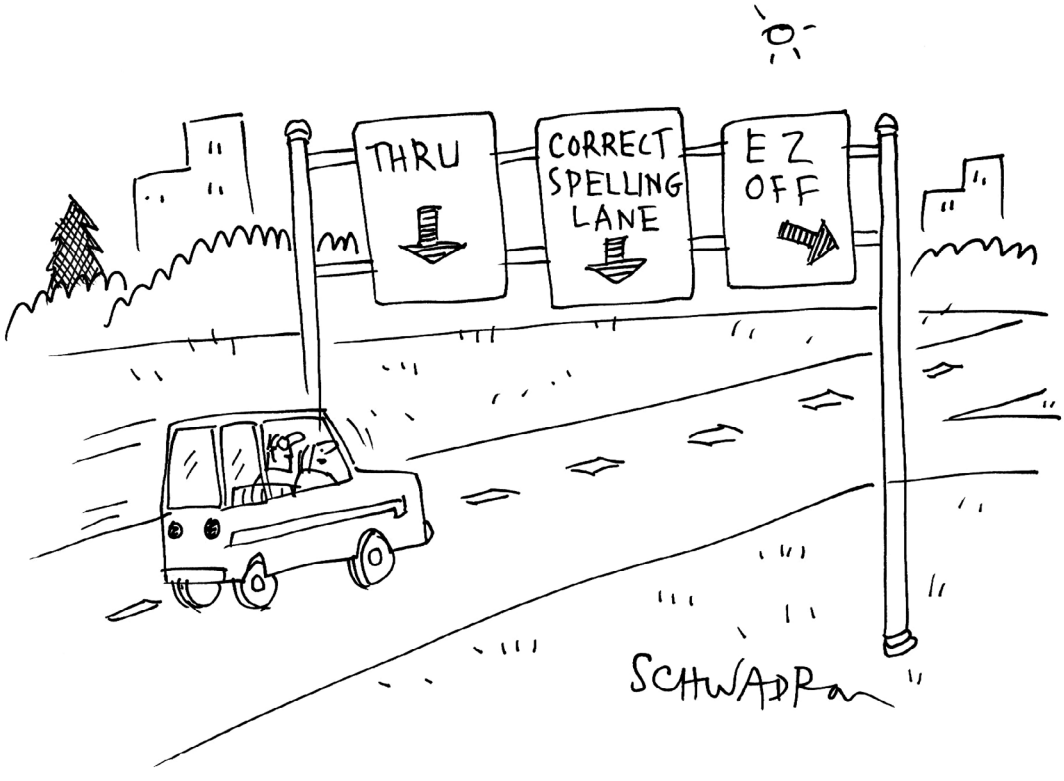
By paying attention to sample sizes, control groups, randomization, and other factors, companies can ensure the validity of their test results. The more valid and repeatable the results, the better they will hold up in the face of internal resistance, which can be especially strong when the results challenge long-standing industry practices and conventional wisdom.

When Petco executives investigated new pricing for a product sold by weight, the results were unequivocal. By far, the best price was for a quarter pound of the product, and that price was for an amount that ended in $.25. That result went sharply against the grain of conventional wisdom, which typically calls for prices ending in 9, such as $4.99 or $2.49. "This broke a rule in retailing that you can't have an 'ugly' price," notes Rhoades. At first, executives at Petco were skeptical of the results, but because the experiment had been conducted so rigorously, they eventually were willing to give the new pricing a try. A targeted rollout confirmed the results, leading to a sales jump of more than 24% after six months.

The lesson is not merely that business experimentation can lead to better ways of doing things. It can also give companies the confidence to overturn wrongheaded conventional wisdom and the faulty business intuition that even seasoned executives can display. And smarter decision making ultimately leads to improved performance.

Could J.C. Penney have averted disaster by rigorously testing the components of its overhaul? At this point, it's impossible to know. But one thing's for certain: Before attempting to implement such a bold program, the company needed to make sure that knowledge—not just intuition—was guiding the decision.

HBR Reprint R1412D



HARLEY SCHWADRON