

How to Design



Managers now have the tools to conduct small-scale tests and gain real insight. But too many “experiments” don’t prove much of anything.

EVERY DAY, managers in your organization take steps to implement new ideas without having any real evidence to back them up. They fiddle with offerings, try out distribution approaches, and alter how work gets done, usually acting on little more than gut feel or seeming common sense – “I’ll bet this” or “I think that.” Even more disturbing, some wrap their decisions in the language of science, creating an illusion of evidence. Their so-called experiments aren’t worthy of the name, because they lack investigative rigor. It’s likely that the resulting guesses will be wrong and, worst of all, that very little will have been learned in the process.

Take the example of a major retail bank that set the goal of improving customer service. It embarked on a program hailed as scientific: Some branches

Smart Business Experiments

by Thomas H. Davenport

Katy Lemay

were labeled “laboratories”; the new approaches being tried were known as “experiments.” Unfortunately, however, the methodology wasn’t as rigorous as the rhetoric implied. Eager to try out a variety of ideas, the bank changed many things at once in its “labs,” making it difficult if not impossible to determine what was really driving any improved results. Branches undergoing interventions weren’t matched to control sites for the most part, so no one could say for sure that the outcomes noted wouldn’t have happened anyway. Anxious to head off criticism, managers did provide a control in one test, which was designed to see if placing video screens showing television news over waiting lines would shorten customers’ perceived waiting time. But rather than looking at control and test groups, they compared just one control site with one test site. That wasn’t enough to ensure statistically valid results. Perceived waiting time did drop in the test branch, but it went up substantially in the control branch, despite no changes there. Those confounding data kept the test from being at all conclusive – but that’s not how the findings were presented to top management.

It doesn’t have to be this way. Thanks to new, broadly available software and given some straightforward investments to build capabilities, managers can now base consequential decisions on scientifically valid experiments. Of course, the scientific method is not new, nor is its application in business. The R&D centers of firms ranging from biscuit bakers to drug makers have always relied on it, as have direct-mail marketers tracking response rates to different permutations of their pitches. To apply it outside such settings, however, has until recently been a major undertaking. Any foray into the randomized testing of management ideas – that is, the random assignment of subjects to test and control groups – meant employing or engaging a PhD in statistics or perhaps a “design of experiments” expert (sometimes seen in advanced TQM pro-

IDEA
IN BRIEF

- » Too many business innovations are launched on a wing and a prayer – despite the fact that it’s now reasonable to expect truly valid tests.
- » With a small investment in training, readily available software, and the right encouragement, an organization can build a “test and learn” capability.
- » Companies that equip managers to perform small-scale yet rigorous experiments don’t only save themselves from expensive mistakes – they also make it more likely that great ideas will see the light of day.

grams). Now, a quantitatively trained MBA can oversee the process, assisted by software that will help determine what kind of samples are necessary, which sites to use for testing and controls, and whether any changes resulting from experiments are statistically significant.

Consumer-facing companies rich in transaction data are already routinely testing innovations well outside the realm of product R&D. They include banks such as PNC, Toronto-Dominion, and Wells Fargo; retailers such as CKE Restaurants, Famous Footwear, Food Lion, Sears, and Subway; and online firms such as Amazon, eBay, and Google. As randomized testing becomes standard procedure in certain settings – website analysis, for instance – firms build the capabilities to apply it in other circumstances as well. (See the sidebar “Stop

Wondering” for a sampling of tests conducted recently.) To be sure, there remain many business situations where it is not easy or practical to structure a scientifically valid experiment. But while the “test and learn” approach might not always be appropriate (no management method is), it will doubtless gain ground over time. Will it do so in your organization? If it’s like many companies I have studied, an investment in software and training will yield quick returns of the low-hanging-fruit variety. The real payoff, however, will happen when the organization as a whole shifts to a test-and-learn mind-set.

When Testing Makes Sense

Formalized testing can provide a level of understanding about what really works that puts more intuitive approaches to shame. In theory, it makes sense for any part of the business in which variation can lead to differential results. In practice, however, there are times when a test is impossible or unnecessary. Some new offerings simply can’t be tested on a small scale. When Best Buy, for example, explored partnering with Paul McCartney on an exclusively marketed CD and a spon-

The real payoff will happen

when the organization as a whole shifts to a test-and-learn mind-set.

sored concert tour, neither component of the promotion could be tested on a small scale, so the company's managers went with their intuition. At Toronto-Dominion, one of the largest and most profitable banks in Canada, testing is so well established that occasionally managers are reminded that, in the interests of speed, they can make the call without a test when they have a great deal of experience in the relevant business domain.

Generally speaking, the triumphs of testing occur in strategy execution, not strategy formulation. Whether in marketing, store or branch location analysis, or website design, the most reliable insights relate to the potential impact and value of tactical changes: a new store format, for example, or marketing promotion or service process. Scientific method is not well suited to assessing a major change in business models, a large merger or acquisition, or some other game-changing decision.

Capital One's experience hints at the natural limits of experimental testing in a business. The company has been one of the world's most aggressive testers since 1988, when its CEO and cofounder, Rich Fairbank, joined its predecessor firm, Signet Bank. You could even say the firm was founded on the concept. One thing that appealed to Fairbank about the credit card industry was its "ability to turn a business into a scientific laboratory where every decision about product design, marketing, channels of communication, credit lines, customer selection, collection policies and cross-selling decisions could be subjected to systematic testing using thousands of experiments."¹ Capital One adopted what Fairbank calls an information-based strategy, and it paid off: The company became the fifth-largest provider of credit cards in the United States.

Yet when it came time to make the largest decision the company had faced in recent years, Capital One's management concluded that testing would not be useful. Realizing that the business would need other sources of capital to remain independent, the team considered acquiring some regional banks in order to transform itself from a monoline credit provider into a full-service bank. The decision was not tested for a couple of important reasons. First, the nature of the opportunity made it imperative to move quickly; no time was available for even a small-scale test. Second, and more

YOU OR SOMEONE on your team is suggesting a change that just might work. But why act on a hunch when you can hold out for evidence? According to the author, the best way to support decision making on potential innovations is to...

» **Design an experiment.**

Start with a hypothesis about how the change will help the business. If it's a good one, you'll learn as much by disproving it as you would by proving it. Put it to the test by measuring what happens in a test group versus a control group. From the outset, be clear on what you need to measure to produce a decisive result – and whether that's a metric you even have the capability to track.

» **Act on the facts.**

Nothing but a success in a testing environment should be rolled out more broadly. But neither should failures simply be scrapped. Refine the hypothesis on the basis of the results, and consider testing a variation. Most important, capture what's been learned, and make it available to others in the organization through a "learning library," so resources aren't wasted proving the same thing again.

EXAMPLE Marketers at the Subway restaurant chain wanted to drum up business by putting foot-long subs on sale for only \$5, but franchise owners worried that the promotion would lure existing customers away from higher-priced menu items. An experiment pitting test sites against control sites proved that the promotion would pay off – which it subsequently did.

» **Make testing the norm.**

Create the training and infrastructure that will enable nonexperts in statistics to oversee rigorous experiments. Off-the-shelf software can walk them through the steps and help them analyze results. A core group of experts can lend resources and expertise and maintain the learning library. Leadership must cultivate a test-and-learn culture, in part by penalizing those who act without sufficient evidence.

As your managers become more comfortable with testing, they'll discover that it paves the way for, rather than throwing up barriers to, promising new ideas.

critical, it was impossible to design an experiment that could reliably predict the outcomes of such a major change in business direction. Still, after making the acquisitions, Capital One reaffirmed its commitment to information-based strategy. Its managers immediately set about translating that ethos into the full-service banking context, which required pushing the method further, into tests involving customer service and employee behavior. As one employee told me, "It's much easier to do randomized testing with direct-mail envelopes than with branch bankers."

Sears Holdings provides another example of what can reasonably be tested and what can't. Interestingly, this is another business with a heritage of testing. Robert E. Wood, who originally moved Sears out of the catalog business and into retail stores, said his favorite book was the *Statistical Abstract of the United States*. When he opened Sears's first free-standing retail stores, in 1928, he placed two in Chicago. Asked why he needed

two in one city, Wood said it was to reduce the risk of choosing a wrong location or store manager.

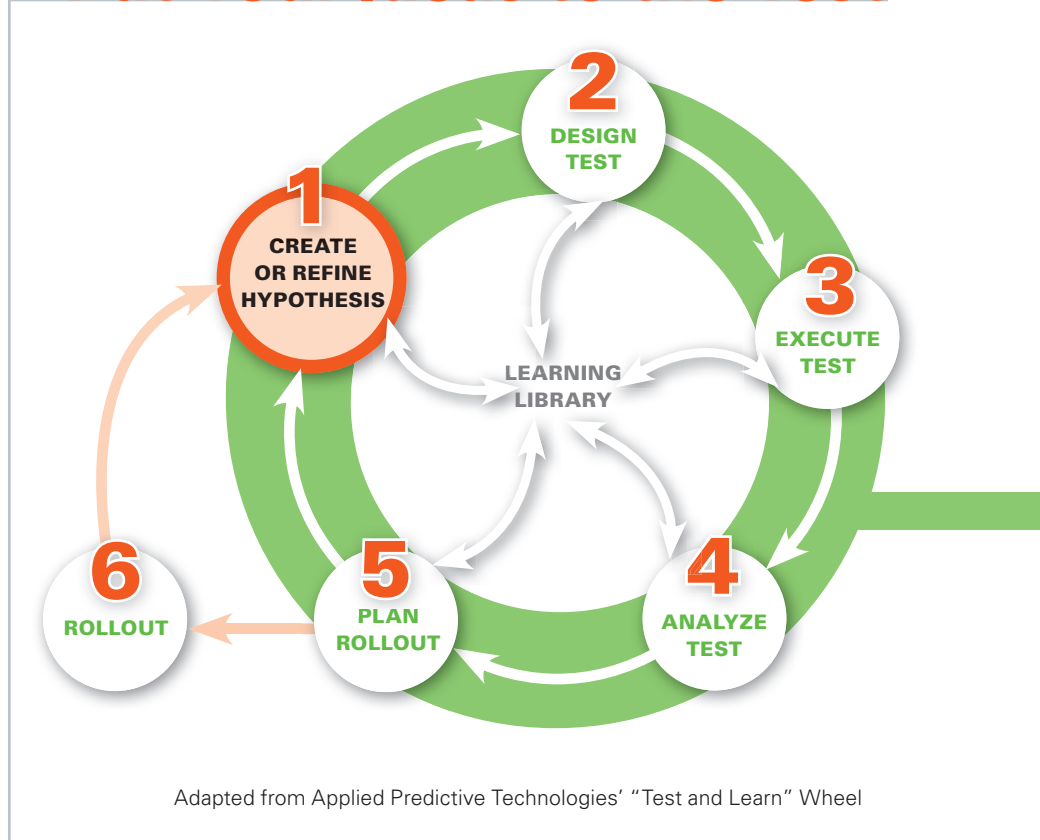
Today Sears Holdings has embarked upon a new era: Its primary owner, financier Edward Lampert, who has been its chairman since Kmart acquired Sears, is exploring alternative ways to combine the two troubled chains. To my knowledge, Lampert didn't test the idea of combining the retailers. That would have been difficult if not impossible to do (and the jury is still out on whether the acquisition was a good decision). However, he's a strong advocate of testing at the tactical level. He wrote in a 2006 letter to shareholders, "One of the great advantages of having approximately 2,300 large-format stores at Sears Holdings is that we can test concepts in a few stores before undertaking the risk and capital associated with rolling out the concept to a larger number of stores or to the entire chain." The retailer has tested, for example, various formats

for including Sears merchandise in Kmart stores, and vice versa, as well as other formats, such as the arrangement of merchandise in Sears stores by rooms in a consumer's home (kitchen, laundry room, bedroom, and so on).

Beyond using the tactical-versus-strategic criterion, there are other ways to decide whether formal testing makes sense. For instance, it is useful only in situations where desired outcomes are defined and measurable. A new sales training program might be proposed, but before you can test its efficacy, you'll need to identify a goal (such as "We want to increase cross-selling"), and you must be able to measure that change (do you even track cross-selling?). Sales and conversion-rate changes are frequently used as dependent variables in tests and are reliably measured for separate purposes. Other outcomes, such as customer satisfaction and employee engagement, may require more effort and invasiveness to measure.

Tests are most reliable where many roughly equivalent settings can be observed. This might mean physical sites, as with Sears's stores, or it might mean more ephemeral settings, such as alternative website versions. Among the earliest and most extensive users of testing are retail and restaurant chains. Because so much is held constant among their multitudinous sites, it is easy to designate which ones will serve as experiments and which will serve as controls and to attribute cause to effect. By the same token, workplace design changes

Put Your Ideas to the Test



Adapted from Applied Predictive Technologies' "Test and Learn" Wheel

are most readily tested in companies that have offices in many cities. Drawing statistical inferences from small numbers of test sites is much more difficult and represents the leading edge of the test-and-learn approach.

Finally, formal testing makes sense only if a logical hypothesis has been formulated about how a proposed intervention will affect a business. Although it's possible to just make a change and then sit back and observe what happens, that process will inevitably lead to a hypothesis – and often the realization that it could have been formulated in advance and tested more precisely.

The Process of Testing

To begin incorporating more scientific management into your business, you'll need to acquaint managers at all levels with your organization's process of testing. It is probably simple to grasp (a typical depiction is shown in the exhibit "Put Your Ideas to the Test"), but it must be communicated in the same terms to people across the organization. Having a shared understanding of what constitutes a valid test enables the innovators to deliver on it and the senior executives to demand it.

The process always begins with the creation of a testable hypothesis. (It should be possible to pass or fail the test based on the measured goals of the hypothesis.) Then the details of the test are designed, which means identifying sites or units

1 CREATE OR REFINE HYPOTHESIS

ASCERTAIN that the hypothesized relationships haven't already been tested and measured – and that they can be.

MAKE sure the hypothesis could generate substantial economic value.

DETERMINE whether it suggests an actual decision or action. (If not, go no further.)

2 EXECUTE TEST

MEET with test and control site managers and analytical experts to discuss what might go wrong and what would constitute test-confounding events.

INSTRUCT field personnel to report abnormal events.

REMOVE sites from test if test-confounding events occur.

ADJUST evaluation and compensation plans for managers so that they are not negatively affected by tests.

3 PLAN ROLLOUT

STUDY attributes of test sites to determine whether rollout should be universal or differentiated.

BALANCE complexity of rollout with ease of implementation and management.

4 LEARNING LIBRARY

DEVELOP a summary of each test: hypotheses, test dimensions, key results, interactions, and rollout strategies and results.

EMPLOY standard business taxonomy to allow easy searching of library.

MAKE library widely accessible to employees; publicize tests and results of important studies to encourage a test-and-learn culture.

1

2

3

4

5

6

DESIGN TEST

ENSURE that the number of test and control sites is sufficient for statistical significance.

USE simulation to explore multiple strategies for creating control groups (for instance, they may be nearly identical but different on one key variable).

ASSESS whether control group strategies previously used for similar tests will suffice; they usually do.

CONDUCT statistical analysis to minimize the number of test cells needed.

EXTEND testing period if key metrics are highly variable.

ANALYZE TEST

ENSURE that "lift" from interventions is statistically significant.

USE software to analyze results and manage complex data from multiple test and control sites.

DETERMINE need for further testing.

EXAMINE as many site attributes as possible to see how key variables interact.

ROLLOUT

STAGGER the rollout and view it as a test in itself. (Are early-adopting sites yielding the desired result? If not, modify the approach in later-adopting sites.)

ENCOURAGE site managers to share rollout strategies and tactics.

to be tested, selecting the control groups, and defining the test and control situations. After the test is carried out for the specified period – which sometimes can take several months but is usually done in less time – the data are analyzed to determine the results and appropriate actions. The results are ideally put into some sort of "learning library" (although, unfortunately, many organizations skip this step). They might lead to a wider rollout of the experiment or further testing of a revised hypothesis.

More broadly, managers must understand how the testing process fits in with other business processes. They conduct tests in the context of, for example, order management, or site selection, or website development, and the testing feeds into various subprocesses. At CKE Restaurants, which includes the

Hardee's and Carl's Jr. quick-service restaurant chains, the process for new product introduction calls for rigorous testing at a certain stage. It starts with brainstorming, in which several cross-functional groups develop a variety of new product ideas. Only some of them make it past the next phase, judgmental screening, during which a group of marketing, product development, and operations people will evaluate ideas based on experience and intuition. Those that make the cut are actually developed and then tested in stores, with well-defined measures and control groups. At that point, executives decide whether to roll out a product systemwide, modify it for retesting, or kill the whole idea.

CKE has attained an enviable hit rate in new product introductions – about one in four new products is successful, versus

one in 50 or 60 for consumer products – and executives say that their rigorous testing process is part of the reason why. If you have had occasion to enjoy a Monster Thickburger at Hardee's, or a Philly Cheesesteak Burger or a Pastrami Burger at Carl's Jr., you've been the beneficiary of CKE's efforts. These are just three of the successful new products that were rolled out after testing proved they would sell well.

At eBay, there is an overarching process for making website changes, and randomized testing is a key component. Like other online businesses, eBay benefits greatly from the fact that it is relatively easy to perform randomized tests of website variations. Its managers have conducted thousands of experiments with different aspects of its website, and because the site garners over a billion page views per day, they are able to conduct multiple experiments concurrently and not run out of treatment and control groups. Simple A/B experiments (comparing two versions of a website) can be structured within a few days, and they typically last at least a week so that they cover full auction periods for selected items. Larger, multivariate experiments may run for more than a month.

Online testing at eBay follows a well-defined process that consists of the following steps:

- Hypothesis development
- Design of the experiment: determining test samples, experimental treatments, and other factors
- Setup of the experiment: assessing costs, determining how to prototype, ensuring fit with the site's performance (for example, making sure the testing doesn't slow down user response time)
- Launch of the experiment: figuring out how long to run it, serving the treatment to users
- Tracking and monitoring
- Analysis and results

The company has also built its own application, called the eBay Experimentation Platform, to lead testers through the process and keep track of what's being tested at what times on what pages.

As with CKE's new product introductions, however, this online testing is only part of the overall change process for eBay's website. Extensive offline testing also takes place, including lab studies, home visits, participatory design sessions, focus groups, and trade-off analysis of website features – all with customers. The company also conducts quantitative visual-

design research and eye-tracking studies as well as diary studies to see how users feel about potential changes. No significant change to the website is made without extensive study and testing. This meticulous process is clearly one reason why eBay is able to introduce most changes with no backlash from its potentially fractious seller community. The online retailer now averages more than 113 million items for sale in more than 50,000 categories at any given time.

eBay performed extensive online and offline testing, for example, in 2007 and 2008, when it changed its page for viewing items on sale. The page had not been redesigned since 2003, and both customers and eBay designers felt it lacked organization, had inadequate photographs of items, and suffered from haphazard item placement and redundant functionality. After going through all the testing steps, eBay adopted a new site design. It posted photos 200% larger than those in the previous design, added a countdown timer for auctions with 24 hours or less to go, made more prominent the item condition and return policy, and included tabs to make shipping and payment fields easier to navigate. It also included new security features to prevent unau-

thorized changes in site content. Each new feature and function was tested independently with control pages. Measures of page views and bid counts suggest that the redesign was very successful.

Building a Testing Capability

Establishing a standard process is the first step toward building an organizational test-and-learn capability, but it isn't sufficient unto itself. Companies that want testing to be a reliable, effective element of their decision making need to create an infrastructure to make that happen. They need training programs to hone competencies, software to structure and analyze the tests, a means of capturing learning, a process for deciding when to repeat tests, and a central organization to provide expert support for all the above.

Managerial training. At the very least, managers should learn what constitutes a randomized test and when to employ it. Capital One, for example, offers a professional education program on testing and experiment design through its internal training function known as Capital One University. One benefit of hosting a program like this, rather than sending managers outside for training, is the greater emphasis on how



the testing connects to upstream and downstream activities in the business.

Test-and-learn software. Some firms, such as Capital One and eBay, have built their own software for managing experiments, but several off-the-shelf options exist – the most common ones being broad statistical packages and analytical tools like SAS. With every passing year, these tools make it more possible for numerate – but not statistically expert – users to conduct truly defensible experiments. Ease of design and analysis has been a particular focus at Applied Predictive Technologies, whose product leads users through the test-and-learn process, keeps track of test and control groups, and provides a repository for findings to be usefully accessed in the future.

Some software tools are tailored to particular problems or industries. Several packaged tools, for example, are available for the analysis of manufacturing-quality experiments. Likewise, highly specialized tools exist for online-usage testing, such as the web analytics software sold by Omniture and WebTrends and the free tools provided by Google Analytics. As of yet, unfortunately, no single software tool can help organizations with all testing types and contexts.

Learning capture. If a firm does a substantial amount of testing, it will generate a substantial amount of learning about what works and what doesn't. Ideally employees throughout the company would share that knowledge and use it to guide future initiatives. But that happens at few organizations. The head of testing at one online firm admitted, "All of that knowledge is in my head, and we'd be in tough shape if I were hit by a bus." One bank executive justified a lack of shared learning, commenting, "We should probably do more, but we've found that people need to learn from doing the test themselves, even if we've done it before many times." People do learn through personal experience, but one would hope that it's not the only possible way.

Some organizations, however, have begun to address the issue. Capital One captures the learning from its thousands of tests in an online knowledge management system and has ex-

Stop Wondering

TESTING is used to make tactical decisions in a range of business settings, from banks to retailers to dot-coms. Here are some questions various companies are examining:

- Do lobster tanks increase lobster sales at Food Lion supermarkets?
- Does a Kmart with a Sears store inside sell more than an all-Kmart format?
- Do eBay users bid higher in auctions when they can pay by credit card?
- What's the optimum number of loose checks for a Wells Fargo ATM to accept?
- Do Subway promotions on low-fat sandwiches increase sandwich sales?
- Does a Famous Footwear store sell fewer shoes when there is a competitor in the same mall?
- Does a Toronto-Dominion branch get significantly more deposits when open 60 hours a week compared with 40?
- Which promotional offers will most efficiently drive checking account acquisition at PNC Bank?

As a result of their testing, these organizations are finding out whether supposedly better ways of doing business are actually better. Once they learn from their tests, they can spread confirmed better practices throughout their business.

perimented with an even more ambitious system that would use such learning to guide product managers as they develop new offerings. Famous Footwear takes a "billboard" approach; for each test, it captures the results in a one-page document, circulates that throughout the organization, and posts it on the wall outside the testing office.

Regular revisiting. One tricky aspect of establishing a long-term testing approach is determining when to retest. There is no way to know for sure when a test has become obsolete; an experienced analyst needs to assess whether enough factors have changed in the environment to make previous results suspect. Famous Footwear executives feel that the retail store location context – their primary application area for testing – changes enough to merit retesting after about a year. Netflix concluded in 2006 that its five-year-old customer tests needed to be redone; the user base had evolved in that time from internet pioneers to mainstream society members. CKE Restaurants has difficulty deciding whether to retest pricing, particularly in times when commodity prices are increasing fast. Ironically, it is human intuition, not testing or analytics, that must be applied to determine the need for retesting.

Core resource group. Most of the firms that do extensive testing have established a small, somewhat centralized organization to supervise it. The group either actually does the testing, as at PNC Bank, Subway, and Famous Footwear, or – if testing is employed throughout the organization – serves as a resource for methodological and statistical questions, as at Capital One. At PNC Bank, the test-

and-learn group (part of the bank's knowledge management function, which reports to Marketing) views the promotion of its own services around the bank as a priority. It tries to build relationships and trust with key executives so that no major initiatives are undertaken without testing. Without a central coordination point, testing methods may not be sufficiently rigorous, and test and control groups across multiple experiments may confound one another. That said, it's not always easy to influence or coordinate testing even when a central group exists.

Creating a Testing Mind-Set

In addition to making the requisite changes in process, technology, and infrastructure, organizations also need to establish a testing culture. Testing costs money (though not as much as widespread rollouts of new tactics that don't work), and it takes time. Senior managers have to become accustomed to, and even passionate about, the idea that no major change in tactics should be adopted without being tested by people who understand testing.


Ask for evidence. CEOs who firmly believe in testing can change their entire organization's perspective on the issue. When people claim that testing has confirmed the wisdom of their idea, have them walk you through the process they used, and demand at least the level of rigor outlined in the exhibit "Put Your Ideas to the Test."

Give it teeth. Gary Loveman at Harrah's Entertainment has said that "not using a control group" is sufficient rationale for termination at the company. Jeff Bezos of Amazon reportedly fired a group of web designers for changing the website without testing. Toronto-Dominion has a culture in which managers insist on tests for every major initiative involving customers or branches. The CEO, Ed Clark, is a PhD economist who once noted that although the bank might not be perfect, "nobody ever criticizes us for not running the numbers."

Sponsor tests yourself. The best management teams in this regard have institutionalized the process of doing and

reviewing tests. At Famous Footwear, Joe Wood and his senior management team meet with the testing head every two weeks to discuss past tests, upcoming tests, and preliminary and final results. Wood says that the company has made testing a part of management's dialogue and the organization's culture.

...

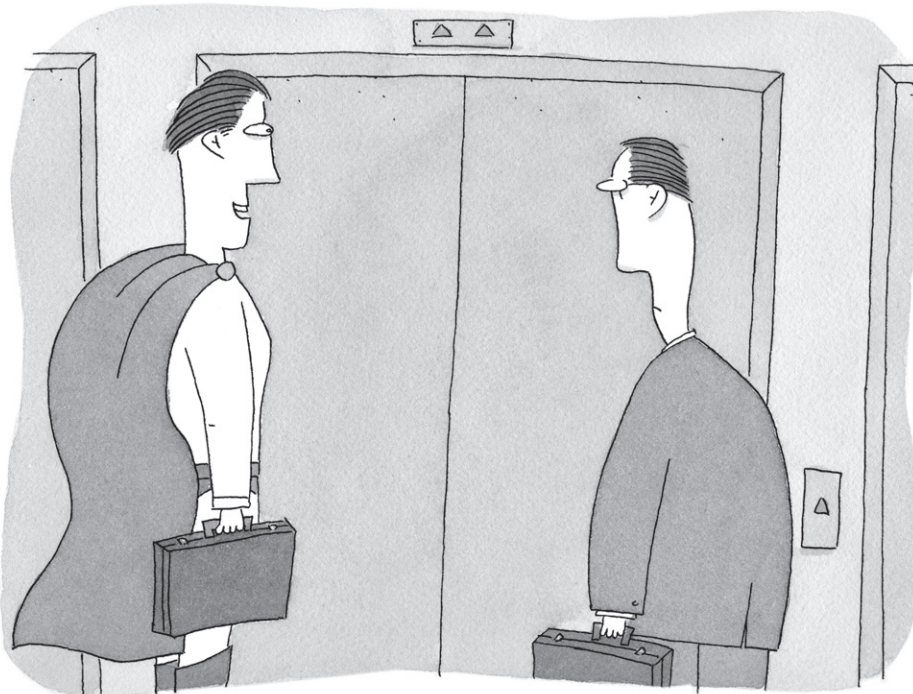
Testing may not be appropriate for every business initiative, but it works for most tactical endeavors. And it just isn't that difficult anymore. It needs to come out of the laboratory and into the boardroom. The key challenges are no longer technological or analytical; they have more to do with simply making managers familiar with the concepts and the process. Testing, and learning from testing, should become central to any organization's decision making. The principles of the scientific method work as well in business as in any other sector of life. It's time to replace "I'll bet" with "I know." 

1. "Capital One Financial Corporation," HBS case no. 9-700-124.

Thomas H. Davenport (tdavenport@babson.edu) is the President's Distinguished Professor of Information Technology and Management at Babson College in Babson Park, Massachusetts. His newest book is *Competing on Analytics: The New Science of Winning*, with Jeanne G. Harris (Harvard Business Press, 2007).

Reprint R0902E

To order, see page 111.



"I'm here to restore confidence in the unrealistic expectations we all had."

Harvard Business Review and Harvard Business School Publishing content on EBSCOhost is licensed for the individual use of authorized EBSCOhost patrons at this institution and is not intended for use as assigned course material. Harvard Business School Publishing is pleased to grant permission to make this work available through "electronic reserves" or other means of digital access or transmission to students enrolled in a course. For rates and authorization regarding such course usage, contact permissions@hbsp.harvard.edu