UNIVERSITY of VIRGINIA | DARDEN Business Publishing

UV7302
Jul. 28, 2017

Behavioral Science in the Marketplace

In the business world, there is often a significant amount of weight placed on first impressions, gut feelings, and common sense. How many of us have seen managers who implement new ideas without having any real evidence to back them up? These are managers who act with little more than common sense or a feel for the situation—the managers who affirm that they can make the right hires, launch the right products, and generally make the right decisions with little more than their intuition guiding their judgment. This phenomenon is known as *management by intuition*.

As powerful as intuition is, it does not consistently lead to optimal outcomes. In recent years, studies have shown that overconfidence is positively related to the introduction of risky products,[1] that trading stock based on intuition leads to worse portfolio performance,[2] and that big-data experimentation can lead to a conversion rate up to 13 times better than intuitive, "best practice" marketing strategies.[3]

Although management by intuition could often be the path of least resistance, we suggest that managers could consider experimentation as a better approach. When the word *experimentation* is mentioned, businesses often raise objections ranging from a fear of failure, to care for customers and not wanting to subject them to less-than-ideal situations, to a desire to potentially increase short-term gains.[4]

However, organizations that succumb to these concerns are doing themselves a disservice and missing out on the advantages of experimentation. Experiments allow for quick vetting of ideas in a systematized, tested way, which in turn often saves time. And since the ideas are tested before implementation, the costs of bad decisions can be minimized and even avoided. Experimentation can also lead to increased confidence and agreement in important decisions, as the chosen courses of action have data to back them up.

Here, we offer an experimental framework to guide your way (**Figure 1**). We begin by explaining what a true experiment looks like and describe when it is appropriate to use one. We then walk through how to plan, design, and run an experiment. Then we explain how to analyze experimental data and how to make business decisions using the results. Finally, we discuss how experiments and their results can be sustainable.

[1] Mark Simon and Susan M. Houghton, "The Relationship between Overconfidence and the Introduction of Risky Products: Evidence from a Field Study," *Academy of Management Journal* 46, no. 2 (2003):139–49.

[2] Mark Fenton-O'Creevy, Emma Soane, Nigel Nicholson, and Paul Willman, "Thinking, Feeling and Deciding: The Influence of Emotions on the Decision Making and Performance of Traders," *Journal of Organizational Behavior* 32, no. 8 (2011): 1044–61.

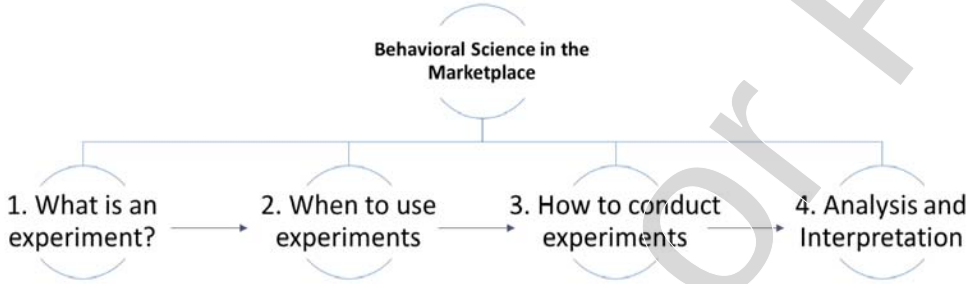
[3] Pål Sundsøy, Johannes Bjelland, Asif M. Iqbal, Alex Pentland, and Yves-Alexandre de Montjoye, "Big Data-Driven Marketing: How Machine Learning Outperforms Marketers' Gut-Feeling," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (Switzerland: Springer International Publishing AG, 2014), 367–74.

[4] Dan Ariely, "Column: Why Businesses Don't Experiment," *Harvard Business Review* 88, no. 4 (2010).

This technical note was prepared by Lalin Anik, Assistant Professor of Business Administration, and Ryan Hauser, MBA Candidate, Yale School of Management.

Experimentation doesn't have to be daunting; by the end of the reading, we hope that you will be an effective designer, analyzer, and proponent of testing who is familiar with the importance (and fun) of experimentation.

Figure 1. Pathway to experimentation.



Source: Created by author.

What is an Experiment?

PRACTICE: Three Problems[5]
Answer the following three questions as quickly as you can: A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

If your first hunch was to answer 10 cents, 100 minutes, or 24 days, congratulations! You have just experienced the power of intuition—specifically, the danger of trusting your gut. It is precisely this sense of (false) confidence that makes management by intuition such an attractive fallback methodology for decision making.[6]

Intuition can be often useful as a shortcut. In psychology, these simple, efficient rules are known as *heuristics* and are integral to efficient decision making.[7] However, these same helpful impressions can also lead us astray. For example, when we use the availability heuristic, we judge the likelihood of an event by the ease with which that event can be brought to mind. This leads us to greatly overestimate the frequency of dramatic, affect-rich events like plane crashes and acts of terrorism. If our intuitive judgments can be so off base in assessing one-line brainteasers like the ones above, how can we expect them to be accurate when the decision in question involves dozens of streams of data, countless probabilistic branches, and overlapping contingencies?

[5] Shane Frederick, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives* 19, no. 4 (2005): 25–42.

[6] The answers are 5 cents, 5 minutes, and 47 days (10 cents, 100 minutes, and 24 days are typical intuitive, though incorrect, answers).

[7] If you want to read more about heuristics, check out Thomas Gilgovich, Dale W. Griffin, and Daniel Kahneman, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge, UK: Cambridge University Press, 2002).

Almost an experiment

Management by intuition, then, is not an optimal practice. Managers are starting to pick up on this, and more companies are increasingly integrating some form of experimentation into their decision-making process. The problem is that many of these experimental processes are quasi-experimental at best. That is, they test some phenomena without using randomization for assigning people to testing groups or using proper controls. For example, pretend Axe seeks to test a new marketing campaign for its body spray. It might opt to test out the new campaign in one city (say, Denver, Colorado) and compare this to the current campaign in another city (say, Kansas City, Missouri). While perhaps cheaper, this quasi-experiment is limited due to the uncontrollable environmental variables in the two cities (economy, competition, and so on). The two cities have several underlying characteristics that make them different that aren't being properly accounted for in this experiment, and these differences may be responsible for causing the groups to react to the campaigns differently. Without randomization and a proper control group, comparisons like this cannot be made.[8]

Additionally, some businesses may rely on focus groups—a handful of people speaking on a subject with which they are not adequately familiar—to set strategies. Others overclaim scientific nomenclature to project an air of proper experimentation (e.g., referring to some locations as "labs" where "experiments" are run), even when this language is not warranted.[9]

While companies are certainly headed in the right direction, it is not enough. Quasi-experiments, focus groups, and using scientific jargon give a false sense of rigor in testing, and can lead to an undeserved comfort in the legitimacy of the findings. If businesses wish to move into a world of experimentation, they need to understand what makes up a true experiment.

Definition of an experiment

An experiment is an investigation in which a hypothesis is scientifically tested, where an independent variable (or *cause*) is manipulated, the dependent variable (or *effect*) is measured, and where extraneous variables are properly controlled for.

In order to conduct a proper experiment and eliminate confounding data, there are certain procedures you need to follow. Understanding these steps involves a small amount of terminology, which we define below. In order to make things more concrete, we will use the same example scenario for each item.

Scenario: Say that Nike wants to test new website landing pages to see which is most effective at causing people to make a purchase on its site. This tactical decision to address the strategy of increasing site purchases lends itself to an experiment.

The main pieces of an experiment are as follows:

- Hypothesis: Your provisional conjecture that will guide your investigation; your starting point for your experiment.

  A good hypothesis is a statement (not a question) that you wish to test; it is your educated guess of something based on what you already know. A good hypothesis is also testable—that is, you need to measure variables associated with the hypothesis to see if it holds or not. Bad hypotheses may involve

[8] John A. List, "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off," *Journal of Economic Perspectives* 25, no. 3 (2011): 3–15.

[9] Thomas H. Davenport, "How to Design Smart Business Experiments," *Harvard Business Review* 87, no. 2 (2009): 68–76.

questions, such as, "Can we increase sales in Russia?" or vague, untestable statements like, "We are the best fashion brand in North America."

While most of the time, we think in terms of confirming data—that is, data that confirms our hypothesis to be true—we must also be receptive to disconfirming data—data that confirms our hypothesis to be false. Even when data are disconfirming and go against your previous assumptions, there is still something to be learned and decisions to inform with this new finding.

Nike might go into the experiment with a hypothesis such as this: "Our new landing page will increase the likelihood of purchase on our website."

- Independent variable(s) (cause): The variable(s) that you are manipulating to test for an effect on the dependent variables.

  The variable that Nike is manipulating is the landing page that a site visitor sees.

- Dependent variable(s) (effect): The variable(s) that you are measuring to test for an effect on the independent variable(s).

  A good dependent variable needs to be both *measurable* and *measured*, two concepts that we will discuss in the section entitled "When to Use Experiments." *Measurable* means that you have a way to measure the variable. For example, revenue, views, and click-through rate are all easily measurable, while things like "creating the most excitement" or "being the best brand" need to be defined more precisely before these variables can be measured. *Measured* simply means that your organization is measuring or has a way to measure the selected variables.

  The variable that Nike is measuring to test for the new landing page's effect is percentage of site visitors who ultimately make a purchase.

- Controlled variables (constants): The factors that must be kept the same between groups (test group and control group) in order to achieve reliable results.

  The factors that Nike must keep the same are the gender, location, and web browser of the visitor, the time of day that a visitor came to the site, as well as others. Nike would do this by randomizing the landing page any given visitor sees. Every visitor, regardless of any personal factor, would have an equally likely chance of seeing the new landing page.

- Test group (Experimental group): The group that receives the manipulation of the independent variable being tested.

  Nike's test group is the visitors seeing the new landing page.

- Control group: The group to which the test group is compared; a baseline group that does not receive the manipulation of the independent variable being tested, used to test the effect of the independent variable. This group represents the natural state of things.

  Nike's control group is the visitors who see the current landing page.

When to Use Experiments

A common misconception exists even among those who know what a true experiment looks like: that you can rely on experimentation across the board for business decisions. This is not the case. There are many different types of experiments and ways of collecting data, each of which applies to a particular setting and may be insufficient on its own. For example, observational data can be limited, big data is often inconclusive, and running multiple regressions provides correlational information but not causal evidence. Additionally, there are instances where any kind of experimentation is inappropriate.

In understanding when to use an experiment, a good place to start is looking at what types of experiments you have at your disposal when making a decision.[10] Below is a list of the most common types of experiments:

Laboratory experiments

These are experiments that are conducted in a well-controlled environment, often (though not necessarily) a laboratory. In these experiments, participants come into the lab (which is often a room with several individual computer stations), are randomly assigned to a group (test versus control), and undergo a standardized procedure developed by the experimenter.

Strengths: The environments of lab experiments allow for precise control of both the independent variable and also extraneous variables. This precision of control allows for easy replication and identification of cause-and-effect relationships.

Weaknesses: The great degree of control could also be its main weakness. That is, since the setting is controlled, the experiment could take on a degree of artificiality, and call into question the generalizability of the findings to the real world (known as ecological validity).

Examples:

- Lab participants indicate their willingness to pay for various displayed items after randomly being told they would have to pay with cash or credit card.
- Lab participants state their liking for a product after being randomly exposed to one of two possible product names.

Online experiments

Many laboratory experiments can also be run online, which is often the quicker and cheaper option. An experiment run online via a subject pool service such as Amazon Mechanical Turk, Qualtrics Panels, Google Consumer Surveys, or Pollfish falls into this category. These types of experiments typically take the form of questionnaires or surveys, but can involve more complex designs.

Strengths: Due to the ease of online distribution and participant recruitment, online experiments are typically fairly cheap and quick. Results can often be obtained in a few days or less.

Weaknesses: There are concerns about the quality of data collected from online participants, as some participants merely aim to finish the task as quickly as possible and earn as much money as they can. This is why things like attention check are crucial (see the section entitled "How to Conduct an Experiment"). Also, as in a lab experiment, the ecological validity of the experiment is questionable.

[10] Gordon L. Patzer, Experiment-Research Methodology in Marketing: Types and Applications (Westport, CT: Quorum Books, 1996).

Examples:

- Online participants on Amazon Mechanical Turk indicate their willingness to pay for various pictured items after randomly being told they would have to pay with cash or credit card.
- Online participants on Pollfish state their liking for a product after being randomly exposed to one of two possible product names.

Field experiments

Field experiments occur exactly where you'd expect—out in the field. These experiments can be natural or manipulated.[11] In a natural experiment, groups are formed by nature or other factors outside the experimenters' control, though the process can still be random. In a manipulated experiment, the experimenter manipulates the independent variable, but the experiment takes place in a natural setting (e.g., a restaurant, a home). These experiments can also be conducted online—for example, on a brand's website.

Strengths: The natural setting of the experiment affords the results with more ecological validity.

Weaknesses: Unlike in the lab, the field experiment might not be very well controlled, which results in less clarity about what variables are causing the results. The better field experiments design controls and randomization as well as collect additional data.

Examples:

- A company tests the effects of sending a $5 coupon to customers in the mail.
- A tea shop studies the impact of giving free samples to its consumers.

A/B testing

A/B testing is found in lab experiments, online experiments, and field experiments, and is popular enough to deserve its own section. In A/B testing, two versions of an item (called A and B) are compared to each other. These items are identical except for one variation which might affect the behavior of some target interacting with the item. Basically, the experimenter is trying to identify one dimension they can change about a message, advertisement, product package, or the like, that will cause the target audience to behave differently.

Examples:

- Visitors to a webpage are randomly assigned to see landing page A or landing page B, and various measures of engagement are measured (e.g., time on page, if a button is clicked or not, likelihood of purchasing a product).
- An e-commerce site randomly assigns site visitors to see one of two color conditions for the "Buy" button, green or yellow. Likelihood of purchase is compared for the two colors.

Big data

While not an experiment per se, we include this methodology due to its overwhelming presence in industry. *Big data* refers to the collection of massive datasets, which are then analyzed to reveal trends, patterns, and associations, especially relating to human behavior and interactions. The main difference between big-data

[11] Duncan Simester, "Field Experiments in Marketing," January 2015.

projects and experiments is the manipulation of an independent variable in a true experiment (and lack thereof in big data).

Big data and experiments can complement each other.[12] One common pairing of the two involves using big data to target and segment the population, and then using experiments to test a desired subgroup.

Examples:

- Barack Obama's 2012 campaign used big data to figure out who the undecided voters were and then experimented with the types of messages they presented to potential voters to see which ones increased votes for Obama the most.
- Insurance agencies use big data to identify individuals who have the highest likelihood of opting into insurance coverage and then experiment with different "nudges" (e.g., defaults and different types of incentives) for different segments to induce them to enroll in a policy.

In experiments, there is often a variable being manipulated, which distinguishes true experiments from simply looking at observational data or big data. Since you manipulate the independent variable of interest, experiments allow you to test for *causal*, rather than just correlational, effects. If you know that A and B are positively related by observing the relationship in observational data, you do not necessarily know if increases in A cause increases in B, increases in B cause increases in A, or if the two just happen to be correlated. With an experiment, if you manipulate A and test B, you can determine if changes in A actually *cause* changes in B. Also, you will notice that to fully take advantage of the experimental types offered above, a business needs to be able to create more than one setting, whether it be multiple locations, different packaging, or multiple iterations of a YouTube advertisement.

Tactics versus strategy and experimentation

The big remaining question of this section is: When should you experiment? To put it simply, when making decisions in a business setting, experiments should be used for tactical decisions, but not strategic ones. While *strategies* define your organization's long-term goals and how you plan to achieve that mission, *tactics* are significantly more focused on steps you are going to take to get there (e.g., specific plans, best practices, resources).

An example of this in terms of general management is one wherein there is an organization that wants to increase profit in Q4. While its strategy could be to cut extraneous costs across the organization, tactics could involve mandating double-sided printing or making personnel cuts.

A marketing example could be one wherein a company wants to generate more traffic to its site. While its strategy is to drive traffic from new, unique visitors to the website, it can use adding website URLs to TV ads or starting 10+ conversations per day with the target audience on Twitter as its tactics.

As you might recognize, strategies represent bigger, more ambiguous goals, while tactics are highly focused actions. It is therefore easy to see why experiments that test strategic decisions are hard or sometimes impossible to implement—there are simply too many confounding variables and contingencies to conduct a proper experiment. For example, think about the complexity involved in trying to predict the outcomes of major changes in business direction (e.g., "Should we change business models?" "Should we proceed with the M&A?"). The outcomes of these decisions are simply too multifaceted to predict and test. Experiments that

[12] Spyros Zoumpoulis, Duncan Simester, and Theos Evgeniou, "Run Field Experiments to Make Sense of Your Big Data," *Harvard Business Review*, November 12, 2015.

test tactical decisions, however, are a lot easier to implement due to the focused nature of tactics. It is easy to imagine testing far less complex tactical decisions (e.g., "To increase sales, what time of the year would be best to offer promotions?" "Which product color is most attractive to our existing customers?"). The outcomes of tactical decisions are typically both *measured* and *measurable*, two topics we will discuss further.

Due to these differences, while experiments can play an important role in decision making, one distinction is crucial to make: experiments are useful in *tactical*, rather than strategic, endeavors.

Even within tactical decisions, research questions can range from the broad to the specific. Ultimately, you want to be able to test as specific a question as possible. A question like, "Is advertising worth the cost?" would be too big to tackle with a simple experiment. The question needs to be focused, so you could narrow it down to something like, "Is advertising during the Super Bowl worth the cost?" or, "How much does advertising our brand name on Google AdWords increase monthly sales?" Similarly, "Should we increase our annual bonuses?" can be improved to something like, "Do year-end bonuses impact the following year's performance?" or "When is the best time to offer financial bonuses?"

One thing to remember is that formal testing only makes sense if a logical hypothesis has been formulated about how a proposed intervention will affect the business. Make sure to vet the hypothesized outcome on the economic value it stands to generate—it should be substantial enough to warrant experimentation. That is, the experiment should be net positive; the benefits resulting from the experiment should exceed the costs.

In sum, experiments can be used to look at the *causal effects* of *tactical*, rather than strategic, business decisions, but only for *specific questions* wherein the results are both *measured and measurable* and are estimated to provide *net-positive results* for the business.

How to Conduct an Experiment

Plan

The first step in conducting an experiment is to identify a hypothesized relationship applicable to your business. To do this, it is often easiest to construct a counterfactual for what your ideal customer behavior looks like. A counterfactual is a statement that pretends that the state of the world (or your business) is different from what it is currently. What is the specific customer behavior you want to change (e.g., customers sign up for a credit card but don't use it)? What is the counterfactual that you want to work toward? You should try to be as specific as possible in this step. A good question to ask yourself in this stage would be something like: "Assume that you are a team of geniuses and everything goes perfectly according to your plan; your customers behave in the exact way that you want them to. Describe what the ideal customer behavior would look like."

To help with this, we have provided you with a behavior-change worksheet in **Exhibit 1** as well as a blank sheet in **Exhibit 2**. Here are the general steps:

1. Describe your target audience and the behaviors its members engage in, with the goal of getting you to think deeply about your target (e.g., who are they, what drives them).
2. Pick a specific behavior you would like to tackle. This helps with simplifying and narrowing down your focus on a behavior rather than an outcome or a feeling.
3. Write out the consumer's behavioral steps, highlighting those they take in interacting with you. This step is similar to drawing out the consumer journey, which will help you figure out where the bottleneck might be in how your consumer interacts with your organization, product, or service. Being able to see

the different steps will help you with exploring at which point you would like to administer an experiment.

4. Think, as you are spelling out the steps, about sources of friction that could be preventing your consumer from engaging in the desired behavior. These sources of friction could be physical, psychological, social, or emotional.

All of these steps will help you with coming up with a hypothesis that will guide your experimentation. Ultimately, this hypothesis should include a *measured* and *measurable* result that you will test. A *measurable* result is one that is quantifiable. For example, sales, click-through rate, and views are all obviously measurable. Some variables, like store atmosphere, are not measurable, but can be made measurable via something like an exit survey (e.g., "From 1 to 7, how pleasant did you find the store's atmosphere?"). *Measured* simply means a piece of data that your organization is keeping track of. You can imagine the tensions that would arise if your experiment tested which ad increased click-through rate, only to find out that your organization doesn't track click-through rate.

Example:

Hypothesis: Changing our website's donation page from an image of a woman to an image of a woman holding a child will increase donations.

Result: The result—increase in donations—is both measurable (it is a numerical value) and measured (your organization is currently keeping track of this value).

Additionally, when you are in the planning stage of the experiment, you need to involve the crucial actors and bring all key decision-makers in as early as possible. If this step is skipped, communication falls by the wayside. Imagine a scenario wherein someone tests the effectiveness of a new button color on a landing page only to find out that the company is under contract with the site designer and cannot change the landing page in any way for two years. Or perhaps someone tests the effect of different e-mail language on customer perception of the e-mail without knowing that the marketing team decided to move away from e-mail communication. Each botched (note: *not* failed) experiment negatively affects the company perception of experimentation, and thus it is key that each test is done carefully and with the involvement of all relevant actors.

Design

When designing the experiment, you need to make sure of a few things:

- Two or more conditions: You need to have both a test group and a control group, as you need a group to use as a baseline for the independent variable's effect.
- Large enough sample size: You need to test enough people to obtain statistically significant results. You know from statistics that a sample size (often called *N*) of 10 will not lead to any significant results.
- Random assignment: In order to mitigate confounding variables and avoid selection bias, assign participants to conditions in your experiment randomly whenever possible. This will help to keep the groups you're testing as equivalent as possible. If you are not able to randomly assign people to certain treatments, another option is to rotate them over time, switching the treatment each group gets. This might help ease the fear of unfairness that some consumers receive different treatment than others. However, this should be done very carefully and only after each wave of the experiment has been completed.

- Hold things constant: The settings for the two groups you are testing should also be as close to equivalent as possible. For example, if you have a sample of 20 stores you wish to test across Texas (10) and New York (10), don't make the 10 Texas stores the test condition and the 10 New York stores the control. The confounding variables of the environment could greatly bias the results. Instead, try to hold extraneous variables like this constant, which you could do in this example by having a test group of 5 Texas stores and 5 New York stores and a control group of 5 Texas stores and 5 New York stores.

- Attention checks: These are used to make sure the participant is engaged with the task and not just randomly answering to finish the experiment as quickly as possible and get paid. If you've got a long series of questions in a row, toss in a "Please select choice 2," or "Are you taking a survey right now?" This way you can clean the data as much as possible and get rid of inattentive noise.

Additionally, you need to consider where you'd like to run your experiment. Will it be in-field (e.g., on your website, across some of your stores)? Will it be online through a panel of participants (e.g., Amazon Mechanical Turk, Qualtrics)?

Run

Once you've properly designed the experiment, you get to launch it. It's exciting to see the results pouring in, but this stage also begets two important questions: (1) When do you stop the experiment? and (2) What do you do if it looks like your experiment is a total failure?

When do you stop the experiment?

As the data come in and you start looking at the preliminary patterns and results, how do you decide when to stop? When do you decide that you have enough data to make the important tactical decisions for your organization?

- One option is to predetermine an end number in terms of how many customers your organization wants to involve (e.g., deciding to run the experiment with 10,000 customers).

- Another option is to dictate a time frame. Note that the time frame you select should be appropriate to address the question you want answered. For example, if the question you're testing involves seasonal differences, you might want to run the study for at least a year.

What if it looks like a dud?

How do you know when to call it quits? If the results are coming in and don't at first blush seem to conform to your hypothesis, should you keep going with the experiment? Should you let it run for another few days or weeks?

If you get no significant difference between the different groups, you should revisit your design. Let's say that you are trying to decrease consumers' electricity consumption. You run an experiment wherein you provide private information to each household about its energy consumption and do not observe any change in consumption patterns in the next three months. This might be disappointing, but there might be important findings from these insignificant results. At that time, you might want to figure out whether you provided the right incentives for consumers to care about energy consumption. Perhaps, in your sample, there are households that do not pay their electricity bills or do not care about private feedback. These same households might care about *public* feedback (e.g., comparisons with other households). You might also think about a reward to motivate behavior or consider framing energy consumption in terms of environmental or health impacts.

Experimentation involves repetition, reconsideration, and revisions. Do not be discouraged if at first the results look unimpressive. One way to mitigate this is to run a pilot test with a smaller sample and shorter duration first to see if you observe any changes. If not, you can always revisit your design for potential improvements.

Analysis and Interpretation

Designing and running the experiment matters very little if the results aren't understood and acted upon appropriately.

Statistical significance

Many managers who run any sort of testing merely look at "lift" from a condition and never bother to test whether the result is statistically significant. While a directional trend may hold some future promise, it could also just be randomness in the data with no further meaning. This is why testing for significance is so crucial.[13]

Scalability

In thinking about how your results might direct your organization tactically, you need to also consider if your experiment will scale or not. This is especially a concern with field experiments. Something like a new landing page for a website will probably scale very well; the page will be uniform regardless of where it is loaded. However, something like a new store policy where customers are greeted by at least three employees may not scale as well; employees may become more lax or policies may not be adhered to as much when the store is not under the strict supervision of the experiment. It does not mean that you should not test the latter design, but be prepared for it to have higher costs (e.g., training, supervision, measurement) in the future.

Precision

While the results from your experiment might be beautiful, you need to ask yourself if they will hold over time. Say you manage a fashion store and have run an experiment in the summer in which you discovered that a new in-store policy of dressing your employees in pastels has increased sales in those stores. This is great news, but you should also be concerned with the precision of the findings. Will this trend still hold in the fall? Will they hold next year once people have seen the employees in pastels for a year straight? Will they hold if a popular blog writes a post deeming pastels uncool?

Additionally, you need to be careful about how much stock you place in the absolute values that come out of an experiment. Every experiment has variation, and thus the absolute values of results will change every time you run another iteration of an experiment. What you *should* be concerned about is the directional relationships that emerge. For example, the finding that a free trial sells better than a discount is a safer result to rely on than that 2.5% of people will respond to a free trial and 1.5% will respond to a discount. Even if the result is significant, the raw percentages will fluctuate, so if contingency plans or estimations are put in place that depend on that raw value, an issue may well arise.

Sustainability

Many field experiments focus on single interactions in the short term as opposed to long-term effects. As a result, the behavioral interventions that influence choices when they are ongoing might be costly if they are

[13] For whatever statistical test you're running, $p < 0.05$ is usually the gold standard for a result being considered significant, though $p < 0.1$ is also sometimes used.

discontinued. There is not yet a magical formula for their sustainability over time. However, one way to extend the effectiveness of experiments and behavioral interventions is to continue running them over a longer period such that people's behaviors start changing in a deeper manner. For example, if you are conducting an experiment on energy use, encouraging consumers to change their lightbulbs will create a physical (technological) difference that might lower the marginal cost of improved decisions over the long term. Since past behaviors and choices often result in more automatic future decisions, where people carry out behaviors with less mental attention, interventions aimed at this short- versus long-term link have higher potential.[14]

Final Thoughts

Companies are increasingly turning to experiments as a way to understand human behavior. Simple, randomized, controlled experiments are guiding the way to uncovering nuances, defining tactics, and impacting business strategy. Your role as a manager is crucial in moving your organization from making decisions by intuition to experimentation. Intuition will still guide your way toward innovation, but it needs to be carefully validated before widespread implementation. The companies of tomorrow that will transform the world are those that take risks, test, learn, and execute. We hope that our guide to behavioral science in the marketplace, while not all encompassing, will help you develop more confidence in the power of experimentation.

[14] Keep in mind that your organization or university most likely has statistical software you can download for free. Absolutely take advantage of this. Excel is a great tool and can be very powerful when used correctly, but there are many operations that are a nightmare in Excel but a breeze in a statistical package. If you don't feel comfortable with statistical programming languages like R or Stata, you can use packages with a clickable user interface (e.g., JMP or SPSS). These packages can almost certainly do everything you'd want to do in testing the results of an experiment in your organization

Exhibit 1

Behavioral Science in the Marketplace

Example Behavior-Change Sheet

Describe Your Target Audience

The more clearly you can describe the people who you are trying to influence, the more easily you can design interventions to encourage the desired behavior.

- Describe the target's demographic characteristics.
- Describe the target's psychographic characteristics: What are their motivations? Dreams, hopes, sources of pride, fears?
- Where and how do they like to shop?

Identify One Specific Behavior

Describe one specific behavior that you want your target to perform. Be sure to describe a specific behavior and not an outcome or feeling.

Not:
- Outcome: Customers engage with your firm digitally.
- Feeling: Customers are more loyal to your firm.

Instead:
- Behavior: Customers create an online account profile when they sign up for a new card.

Describe the Behavioral Steps

Describe the specific steps that your target audience will take to accomplish the desired behavior. The steps may or may not be sequential.

1. Read the e-mail that describes the benefits of digital account access.
2. Click on the link in the e-mail.
3. Complete the desired fields to verify account.
4. Select a username and password.
5. Verify the username and password.
6. Log into new account.

Pinpoint Sources of Friction

Identify the barriers that prevent your target from accomplishing the desired behavior. Identify where and why people might "drop off."

- Cognitive load: People can feel overloaded with too many choices or uncertainty, leading to inaction
- Heuristics: People employ mental shortcuts to "cut through the clutter" and make decisions more easily
- Context: People process information in the context in which it is presented; context affects behavior
- Social cues: Societal norms and peer influence have a major impact on behavior
- Habits: To break an existing habit and create a new one, must disrupt the cycle

Behavioral Step	Friction

Plan Your Test Design

Think about how behavioral principles might address each area of friction. Leveraging these principles, design an experiment that will help you determine how best to achieve the desired behavior.

Points to think about:
- Be specific but simple in your experimental design
- Make sure you have thought about a control group (if applicable) as well as how you will randomly assign customers to each group
- Think about how you will measure success (e.g., a change in a specific behavior)

Source: All exhibits created by author.

Exhibit 2

Behavioral Science in the Marketplace

Blank Behavior-Change Sheet

Describe Your Target Audience	Identify One Specific Behavior	Describe the Behavioral Steps

Pinpoint Sources of Friction	Behavioral Step	Friction	Plan Your Test Design