11 Accelerated Learning by Experimentation

Roger Bohn and Michael A. Lapré

CONTENTS

DELIBERATE LEARNING

Experimentation was a key part of the Scientific Revolution. Galileo (1564–1642) is often credited with being the first to use experiments for both discovery and proof of scientific relationships, although this could also be said of Ibn al-Haytham (965–1040) who lived some 600 years earlier. Experimentation as a core concept in management was introduced by Herbert Simon, who discussed both management and engineering as processes of systematic search over a field of alternatives (March and Simon 1958; Newell and Simon 1972). Deliberate systematic experimentation to improve manufacturing probably goes back to chemical engineering in the late nineteenth century. Frederick Taylor ran thousands of metal-cutting experiments over several decades, and in many ways was a pioneer in systematic learning, a decade before his controversial research on managing people (Bohn 2005). Systematic experimentation in marketing began around 1940 (Applebaum and Spears 1950).

A general definition of an experiment is: "A deliberate comparison of outcomes from a varied but repeatable set of conditions, with an effort to explain different outcomes by differences in conditions." This definition includes controlled experiments (in which possible causal conditions are deliberately manipulated), and natural experiments (in which they are measured but not deliberately altered). The definition excludes purely descriptive investigations such as surveys or satellite photographs in which the results are solely tabulated or displayed. For example, a control chart by itself is not an experiment, although control charts can provide data for natural experiments.

Experiments are a key mechanism for industrial learning, and are therefore an important managerial lever to accelerate learning curves (Adler and Clark 1991; Dutton and Thomas 1994; Lapré et al. 2000). The learning curve phenomenon has been observed frequently for measures of organizational performance such as quality and productivity. The rate of improvement is called the "learning rate." Learning rates show tremendous variation across industries, organizations, and organizational units (Dutton and Thomas 1984; Lapré and Van Wassenhove 2001). Bohn (1994) and Lapré (2011) discuss the model inside the learning curve. Both experience and deliberate activities can be sources for learning; learning can yield better organizational knowledge, which in turn can lead to changed behavior, and subsequently to improved performance. None of these steps are automatic. Dutton and Thomas (1984) call learning from experience "autonomous learning" and learning from deliberate activities "induced learning." The typical examples of deliberate activities are quality improvement projects and productivity improvement projects. Such projects often rely on a series of experiments. Other induced learning methods at the level of the firm are deliberate knowledge transfers from outside the organization, and the training of workers—but these are available only for knowledge that already exists and is accessible. Even when firms transfer knowledge from the outside, some adaptation to local circumstances—and thus experimentation—is almost always required (Leonard-Barton 1988). Hence, sound management of experimentation is important in order to attain the effective management of the learning rate.

In an extreme example, experimentation can be the sole driver of a learning curve. Lapré and Van Wassenhove (2001) studied productivity learning curves at four production lines at Bekaert, the world's largest independent producer of steel wire. One production line was run as a learning laboratory in a factory. The productivity learning curve was explained by the cumulative number of productivity improvement projects, which consisted of a series of natural and controlled experiments. The other three production lines were set up to replicate the induced learning. Interestingly, the other three lines struggled with learning from experimentation and relied on autonomous learning instead. Even within the same organization, it can be difficult to manage the required scientific understanding and human creativity (we will later refer to this as the "value of the underlying ideas").

Experiments are used in a variety of settings. Generally, the experiments themselves are embedded into broader processes for deliberate learning, such as line start-ups, quality programs, product development, or market research. Examples of situations dealt with by experimentation include:

- Diagnosing and solving a newly observed problem in a complex machine
- Improving the layout or contents of web pages that are displayed to consumers
- Breeding new varieties of plants
- Developing a new product by building and testing prototypes
- Conducting a clinical trial of a new medication on a population of patients
- Scientific modeling via simulation, such as global climate change models

Experimentation is not the only method of performance improvement in industry. Other approaches start from existing knowledge and attempt to employ it more widely or more effectively. These include training workers, installing improved machinery (with the knowledge embedded in the machines), and improving process monitoring to respond faster to deviations. In such approaches, experiments are still needed to validate changes, but they do not play a central role.

Sequential Experimentation

Experiments are generally conducted in a series, rather than individually. Each cycle in the series consists of planning the experiment, setting up the experimental apparatus, actually running the trial and collecting data, and analyzing the results (plan, set-up, run, analyze). For example, in product development, experiments focus on prototypes, and the prototyping cycle consists of designing (the next prototype), building, testing (running trials with the prototype), and analyzing (e.g., see Thomke 1998). The "planning" stage includes deciding what specific topic to investigate, deciding where and how to experiment (discussed below), and the detailed design of the experiment. The analysis of results from each experiment in the series suggests further ideas to explore.

Experiments can be designed for the general exploration of a situation, to compare discrete alternative actions, to estimate the coefficients of a mathematical model that will be used for optimization or decision making, or to test a hypothesis about causal relationships in a complex system. The goals of experiments in a series usually evolve over time, such as moving from general to specific knowledge targets. For example, to fix a novel problem in a multistage manufacturing process, engineers may first isolate the location of the cause (general search), and then test hypotheses about the specific causal mechanism at work. They may then test possible interventions to find out whether or not they fix the problem and what, if any, are the side effects. Finally, they may run a series of experiments to quantitatively optimize the solution.

The goals of experiments depend on how much is already known. Knowledge about a particular phenomenon is not a continuous variable, but rather it passes through a series of discrete stages. The current stage of knowledge determines the kinds of experiments needed in order to move to the next stage (Table 11.1). For example, to reach Stage 3 of causal knowledge, one must learn the direction of an effect, which may only require a simple two-level experiment, while at Stage 4 the full effects are understood quantitatively. This requires an elaborate design, often

TABLE 11.1
Stages of Causal Knowledge and the Types of Experiments Needed to Progress

Causal Knowledge Stage: How Cause x_i Affects Outcome Y	Experimental Design Needed to Get to this Stage
6. Integrated multivariate causal system	Multivariate experiment with initial, intermediate, and final variables all measured
5. Scientific model (formulaic relationship derived from scientific theory)	Test fit to an equation derived from first principles
4. Magnitude and shape of effect (empirical relationship)	Compare range of levels of x_i (response surface estimation)
3. Direction of effect on Y	Compare two selected alternatives
2. Awareness of variable x_i	Screening experiment on multiple possible causes; general exploration
1. Ignorance	(Starting condition)

involving the changing of multiple variables. Whether the additional knowledge is worth the additional effort depends on the economics of the particular situation. Moving from rudimentary to complete knowledge often takes years because advanced technologies have complex networks of relationships, with hundreds of variables to consider.

Learning in Semiconductor Manufacturing

Deliberate learning takes place almost constantly in semiconductor fabrication. Fixed costs of fabrication are very high, so the production rate (throughput) and yield (fraction of output that is good) are critical. Yields of new processes sometimes start well below 50%, and a percentage point of yield improvement is worth millions of dollars per month (Weber 2004).

Because of the very high complexity of the fabrication process, no changes are made unless they have been tested experimentally. Hundreds of engineers in each wafer fabrication facility (called "fab") engage in constant learning cycles. Some seek changes that will permanently improve yields. Learning targets can include the product design at several levels, from circuit design to specific masks, changes in the process recipe, changes in the tooling recipe (such as a time/intensity profile for a deposition process), or alterations in a $10 million tool. Other problem-solving activities can locate and diagnose temporary problems, such as contamination. Natural experiments (discussed below) may be adequate to find such problems, but engineers always test the solution by using a controlled experiment before putting the tool back into production. Other experiments are used to test possible product improvements. Finally, some experiments are run for development purposes, either of new products or of new processes.

Learning curves for yield improvement can differ considerably, even within the same company. In the case of one semiconductor company for which we had data, the cumulative production that was required in order to raise yields by 10 percentage points from their initial level ranged from below 10 units, to more than 200 units.[*] There were many reasons behind this large range, including better systems put in place for experimentation and observation (natural experiments), transfer of knowledge across products, and more resources for experiments at some times than at others.

METHODS OF ACCELERATING LEARNING

Deliberate (induced) learning is a managed process. This section discusses the drivers of effective learning by experimentation. There is no perfect way to experiment, and choosing methods involves multiple tradeoffs. In some situations, a better strategy can have a 10-fold effect on the rate of learning. Even when using a single strategy, effectiveness can vary dramatically.

We divide the analysis into three sections. First we show that there are four basic types of experiments. Second, we present criteria for predicting the rate of learning from experiments. These include cost per experiment, speed, and specific statistical properties. Third, we discuss the choice of where to experiment (location). At one extreme are full-scale experiments in the real world, while more controlled locations are usally superior. Different combinations of experiment type and location can improve some of the criteria, while worsening others.

This three-part framework (types, criteria, and locations) was first developed for manufacturing (Bohn 1987) and was then applied to product development (Thomke 1998). The framework also fits market research and other industrial learning about behavior. It can also be used in clinical trials.

Types of Experiments and their Characteristics

There are four main types of experiments, each with different ways of manipulating the causal variables. (1) *Controlled experiments* make deliberate changes to treatments for several groups of subjects, and compare their properties. For example, medical *clinical trials* treat separate groups of patients with different drugs or doses. The "control group" captures the effects of unobserved variables; the difference in outcomes between the control and treated groups is the estimated effect of the treatment. Treatments in controlled experiments can be elaborated indefinitely. A classic reference on this subject is Box et al. (1978).

(2) *Natural experiments* use normal production as the data source.[†] The natural variation in causal variables is measured carefully, and is related to the natural variation in outcomes using regression or related techniques. Natural experiments are generally inexpensive; the main costs involved are for analysis, and, if necessary, special measurements. As a result, very large sample sizes are possible. On the other hand, natural experiments can only measure the impact of changes that occur due to

[*] Arbitrary production units used for disguise.

[†] The term "natural experiment" has apparently never been formally defined, although various authors have used it, or have discussed a similar concept using a different name. Murnane and Nelson (1984) refer to natural experiments but without defining them. Box et al. (1978) referred to them as "happenstance data."

natural variations in the process. They cannot predict the effects of radical changes, such as a new type of equipment, or a completely new procedure.

A fundamental problem with natural experiments is confounding. If A and B vary together, does A cause B, or does B cause A, or are both caused by an invisible third variable (Box et al. 1978)? There are also tricky questions about causality. Suppose that the length of time spent by customers in a grocery store is measured and turns out to be positively correlated to the amount of money spent. Do the customers: (a) spend more because they had more time to look at merchandise, or (b) spend more time shopping because they intended to buy more at the time they entered the store? Gathering additional data via a questionnaire might resolve this, while still remaining a natural experiment, but even if the first case is correct, an intervention that increases the time spent in the store will not necessarily lead to increased spending.[*] This simple example highlights the importance of an explicit *causal model* for learning from experiments; an appropriately complex causal model is needed in order to understand the effect of interventions that can change A, compared with just statistically establishing and concluding that A increases B (Pearl 2001).[†] The causal model can be determined from outside knowledge, or by appropriate controlled experiments, but it cannot generally be tested purely by natural experiments.

(3) *Ad hoc experiments*, like controlled experiments, use deliberate changes. However, the changes are made without a careful control group or experimental design. A simple "before-and-after" comparison is used to estimate the impact of the treatment. Because many unobserved variables can also change over time, ad hoc experiments can be very misleading and can have a poor reputation. However, young children playing with blocks learn very effectively in this way, and quickly learn basic cause-and-effect relationships. This form of learning is sometimes called *trial and error* (Nelson 2008).

(4) *Evolutionary operation (EVOP) experiments* are a hybrid between controlled and natural experiments. Slight changes in the production process are made deliberately, and the resulting changes are measured and statistically associated with the process shifts. The changes are small enough so that the process still works and the output is still good. Subsequent changes can move further on in whichever directions yield the most improved results—this is the "evolution."

EVOP was proposed decades ago for factories, but as far as we know, it was little used (Hunter and Kittrell 1966).[‡] Recently, however, it has become a very common approach to learning on the Internet. As is discussed below, Amazon and Google both use multiple EVOP experiments to tune their user-interface (web page) design, search algorithms for responding to queries, selection and placement of ads on the page, and so forth. Seemingly minor design issues such as colors and the precise location of "hot spots" on the page can be experimented on very easily, quickly, and cheaply.

[*] For example, management could slow down the checkout process, or show a free movie. These might seem far fetched in a grocery store, but similar problems would cloud the results of a natural experiment in a casino, a theme park, or a bookstore.

[†] The standard statistical and mathematical notations are not even capable of distinguishing among the different types of causes (see Pearl 2001).

[‡] It was also proposed for marketing, as a form of "adaptive control" (Little 1966).

All four types of experiments are generally feasible for learning in ongoing operations, but only controlled and ad hoc experiments are available for novel situations. Deciding which type to use depends on the interactions among several criteria, which we turn to next.

Criteria for Evaluating Experimentation Methods

There are advantages and drawbacks to the different types of experiments and the different designs within a type. There is, as yet, no "theory of optimal experimentation," except within very stylized models. However, a small number of high-level properties are essential to predicting the overall effectiveness of the different approaches. Key characteristics include speed, signal-to-noise (S/N) ratio, cost per cycle, value and variety of ideas, and fidelity.

- *Speed* can be defined as the inverse of the *information cycle time* from the beginning to the end of each plan–set-up–run–analyze cycle. Shorter cycle times directly accelerate learning by requiring less time to run a series of experiments. Less directly, a faster cycle helps the experimenters "keep track of" the experiment, the reasoning behind it, and any unrecorded subtleties in the experimental conditions, along with any results that may be important only in retrospect.
- *Signal-to-noise ratio* (S/N) can be defined as the ratio of the true (unknown) effect of the experimental change to the standard deviation of measured outcomes. The S/N drives the common significance test such as the "t test," which measures the probability of a false positive result. It also drives the statistical *power*, which measures the probability of a false negative (overlooking a genuine improvement) (Bohn 1995a). The S/N can generally be improved by increasing the sample size, but more subtle methods are usually available and are often less expensive.[*]
- *Cost* per cycle. The lower the cost, the more experimental cycles can be run (or the more conditions can be tested in parallel). The financial costs of controlled experiments usually include the cost of the materials used in the experiment, but the most important costs are often non-financial—notably, the opportunity costs of busy engineers, the production capacity, computer time (for simulation), or the laboratory-quality metrology equipment. Controlled experiments carried out on production equipment often require elaborate set-ups, which increases the opportunity costs. One of the great benefits of natural experiments and EVOP is that the only costs are for special measurements, if any, and the analysis of the results. Costs are divisible into variable costs (meaning proportional to the sample size) and fixed costs (which depend on the complexity of the experimental set-up, but not the sample size). A third type of cost is capital costs for expenditures to create the experimental system itself. The cost of a pilot line can be considered a capital cost to enable more/better experiments.

[*] S/N is a core concept in communications engineering.

In semiconductor fabs, experimental lots are interspersed with normal production lots and the cost of experimentation is managed by quotas, rather than a financial budget. So-called "hot" lots are accelerated through the process by putting them at the front of the queue at each process step, giving roughly a two-fold reduction in the information cycle time, but increasing the cycle time for normal lots. For example, a development project could be given a quota of five normal lots and two hot lots per week. It is up to the engineers to assign these lots and their wafers to different questions within the overall development effort. Even for hot lots the information cycle is generally more than a month, so one fab could have 50 experiments in progress at one time.

- *Value and variety of the underlying ideas* being tested. Ideas for experiments come from a variety of sources, including prior experiments, outside organizations, scientific understanding of the problem, and human creativity. Strong ideas convey a double benefit: they improve the S/N of the experiment and, if the experiment reaches the correct conclusion, they increase the benefit derived from the new knowledge.

  In mature production processes and markets, most ideas will have a negative expected value—they make the situation worse. A higher variety of underlying ideas raises the total value of experimentation. This follows from thinking of experiments as real options, where the cost of the experiment is the price of buying the option, the current value of "best practice" is its exercise price, and the revealed value of the new method is the value of the underlying asset.* According to the Black-Scholes formula and its variants, higher variance of the asset increases the value of options. There is a further benefit from the higher S/N that is not captured in the standard formulas, namely, the reduced probabilities of statistical errors.
- *Fidelity* of the experiment can be defined as the degree to which the experimental conditions emulate the world in which the results will be applied. Fidelity is a major concern in choosing both the type and the location of experiments, which we discuss later.

Ideal experimentation strategies and tactics would increase the value of all five criteria. More commonly though, the choices of how to experiment will involve tradeoffs between the criteria. For example, S/N ratio increases with sample size, but so does the cost (except for natural experiments). Speed can usually be increased by paying a higher price. The degree to which this is worthwhile depends on the opportunity cost of slower learning. This depends on both business and technical factors. Terwiesch and Bohn (2001) show that under some conditions of factory ramp-up, the optimal policy is bang-bang: at first, carry out zero production and devote all resources to experiments; later carry out no experiments and devote all resources to production. At least in semiconductors, the normal pattern for debugging a new process is to start with 100%

* Terwiesch and Ulrich (2009) show how higher variance is good in *tournaments*. Tournaments are a highly structured form of experimentation, in which only one proposal will be selected out of many.

experiments, then shift to a quota, such as 5% experiments, and go to zero experiments near the end of life.

The recent popularity of "data mining" reflects the power of natural experiments under the right circumstances. For example, Harrah's Casino uses natural experiments for insights that drive a variety of controlled experiments (Lee et al. 2003). In data mining, a company has invested in a database of historical data and analytic hardware, software, and expertise. Once this large investment is operational, each experiment costs only a few hours, or days, of people's time and server processing time. The sample size can be in the millions, so the S/N ratio may be excellent, even if the environment has high variability and the data has measurement errors. So, three of the five criteria are extremely good: S/N, speed, and cost. On the other hand, fidelity is unclear, since much of the data are old and subject to the causal ambiguity problems discussed earlier.

Approximate Reality: The Location of Experiments

The goal of experimentation in industry is to develop knowledge that can be applied in real environments, such as a high-volume manufacturing plant, a network of service delivery points (automated teller machines, web servers, stores, etc.), the end-use environments of a product, a population of customers, or sufferers from a particular disease. Yet it is usually preferable to do experiments elsewhere—in an environment that emulates key characteristics of the real world, but suppresses other characteristics in order to improve the S/N, cost, or information cycle time. Scientists have experimented on models of the real world for centuries, and new product developers have used models at least since the Wright brothers' wind tunnels. The literature on product development identifies two orthogonal choices for how to model: *analytical to physical* and *focused to comprehensive* (Ulrich and Eppinger 2008). Both of these axes apply to all learning domains.[*]

Organizations that do a lot of learning by experimentation often designate special facilities for the purpose, referred to as *pilot lines*, *model stores*, *laboratories*, *test chambers*, *beta software versions*, and other names. These facilities usually have several special characteristics. First, they are more carefully controlled environments than the real world.[†] For example, clinical trials use prescreened subjects, with clear clinical symptoms, no other diseases, and other characteristics that increase the likely signal size and decrease the noise level. Second, they are usually better instrumented, with more variables measured, more accurately and more often. For example, test stores may use video cameras to study shopper behavior. Third, these facilities are more flexible, with more skilled workers and different tools.

Moving even further away from the real world are virtual environments, such as simulation models and mathematical equations. Finite element models have revolutionized experimentation in a number of engineering disciplines, because they allow science-derived first principles to be applied to complex systems.

[*] The original literature on manufacturing experimentation conflated them into a single axis, "location" (Bohn 1987).

[†] But see the Taguchi argument for using normal quality materials, discussed in the conclusion.

Ulrich and Eppinger (2008) call the degree of reality or abstraction the "analytical/physical" dimension. They apply it to prototypes in product development, but their spectrum from physical to analytical also applies in manufacturing, medicine, and other domains, as illustrated in Table 11.2. For example, a steel company found that water and molten steel had about the same flow characteristics, and therefore prototyped new equipment configurations using scale models and water (Leonard-Barton 1995). In drug development, test tube and animal experiments precede human trials.

There are many benefits of moving away from full reality toward analytical approximations (top to bottom in Table 11.2). Cost per experiment, information cycle time, and the S/N ratio all generally improve. The S/N ratio can become very high in deterministic mathematical models. The disadvantage of moving toward analytical models is the loss of fidelity: the results of the experiment will be only partly relevant to the real world. In principle, the best approach is to run an experiment in the simplest possible (most analytical) environment that will still capture the essential elements of the problem under study. In practice, there is no single level of abstraction that is correct for every part of a problem, and learning therefore requires a variety of locations, with preliminary results developed at more analytical levels and then checked in more physical environments.

Full-scale manufacturing can be very complex. Lapré et al. (2000) studied organizational learning efforts at Bekaert, the world's largest independent producer of steel wire. Bekaert's production process can be characterized by detail complexity

TABLE 11.2
The Spectrum of Locations from Physical to Analytical

Locations of Experiments: Physical/Analytic Range	Manufacturing	Aviation Product Development Example	Drug Testing Example	Retail Behavior Example
Full-scale reality (most physical)	Manufacturing line	Flight test	Human trial	Full-scale trial
Scale model	Pilot line	Wind tunnel	Animal model	Test market
Laboratory	Lab		In vitro test	Focus group
Complex mathematical model	Finite-element simulation	CAD (finite element) simulation with graphical output	Rational drug design model	
Simple mathematical model (most analytical)	Simultaneous equation model for annealing in a furnace	Strength of materials model	Half-life model of drug metabolism	Advertising response model

Note: Example locations shown for four learning situations. Fidelity is highest at the top (full scale), but information cycle times, signal-to-noise ratio, and cost per experiment get better as the domain moves toward analytical (at the bottom).

(hundreds of machines, and hundreds of process settings), dynamic complexity (lots of dependencies between the production stages), and incomplete knowledge concerning the relevant process variables and their interactions. They found that Bekaert factories sometimes used the results from experiments run in laboratories at a central research and development (R&D) facility. However, on average, these laboratory insights actually slowed down the rate of learning in full-scale manufacturing. Small-scale laboratories at the central R&D facility were too different from the reality of full-scale manufacturing. Ignoring the complexity of full-scale manufacturing (such as equipment configurations) actually caused deterioration in performance. Thus, in manufacturing environments such as Bekaert's, fidelity issues mean that the locations used for experiments need to be more physical and less analytical. Bekaert, therefore, did most of its experiments in a special "learning line" set up inside its home factory.

Focused/Comprehensive Spectrum

The second aspect of experimental "location" is what Ulrich and Eppinger (2008) call "focused/comprehensive" dimension. An experiment can be run on a subsystem of the entire process or product rather than the entire system. It is easier to try out different automobile door designs by experimenting on a door than it is on an entire automobile, and easier still to experiment on a door latch. The effects of focused studies are similar to the effect of moving from physical to analytical locations. Once again, the tradeoff is loss of fidelity: subsystems have interactions with the rest of the system that will be missed. So this technique is more applicable in highly modular systems. At Bekaert, there were high interactions among the four main process stages, so single-stage trials were risky.

Experimenting on a subsystem has always been common in product development. In manufacturing and other complex processes, it is more difficult and more subtle, yet can have an order-of-magnitude effect on the speed of learning. The key is to understand the process well enough to measure variables that *predict* the final outcome before it actually happens. The case study "short-loop experiments for AIDS" in the next section gives a dramatic example where the learning cycle time was reduced from years to months.

In semiconductor fabrication, experiments that only go through part of the process, such as a single tool or single layer, are sometimes called *short-loop experiments* (Bohn 1995b).[*] Suppose that a megahertz-reducing mechanism has been tentatively diagnosed as occurring in layer 10 out of 30, in a process with an average cycle time of two days per layer. Mathematical models of the product/process interaction suggest a solution, which will be tested by a split-lot experiment in which 14 wafers are treated by the proposed new method and 10 by the old method. Clearly, 18 days can be saved by splitting a previously routine production lot at the end of layer 9, rather than at the beginning of the line. However, must the test lot be processed all the way to the end before measuring the yields of each wafer? If so, the information cycle time will be 42 days, plus a few more for setting up the experiment and testing.

[*] Juran referred to such experiments as "cutting a new window" into the process (Juran and Gryna 1988).

Furthermore, megahertz differences among the 24 wafers will be due to the effect of the process change plus all the normal wafer-to-wafer variation in all 30 layers, leading to a poor S/N ratio (Bohn 1995b).

The information cycle time and S/N will be far better if the effect of the change can be measured directly after the tenth layer, without processing the wafers further. This requires a good understanding of the intermediate cause of the problem, and the ability to measure it accurately in a single layer. Furthermore, it requires confidence that the proposed change will not interact with any later process steps.

Good (fast and accurate) measurement of key properties is critical to focused experiments. To allow better short-loop experiments, most semiconductor designs include special test structures for measuring the electrical properties of key layers and features. The electrical properties of these test structures can be measured and compared across all wafers. Running the trial and measuring the test structure properties can be done in a few days. Depending on how well the problem and solution are understood, this may be sufficient time to go ahead and make a process change. Even then, engineers will need to pay special attention to the megahertz tests of the first production lots using the new method. This checks the fidelity of the early short-loop results against the full effect of the change. If there is a discrepancy, it means that the model relating the test structures to performance of the full device needs to be revised.

An analogous measurement for focused consumer behavior experiments is the ability to measure the emotional effect of advertisements on consumers in real time, and the ability to prove that the measured emotional state predicts their later purchasing behavior. Once these conditions exist, consumers can be shown a variety of advertisements quickly, with little need to measure their actual purchasing behavior. Measuring emotional response using functional magnetic resonance imaging (fMRI) is still in its infancy, but as it becomes easier it will have a big effect on the rate of progress in consumer marketing situations (for a review on fMRI, see Logothetis 2008).

CASE STUDIES

This section illustrates how learning has been accelerated through more effective experimentation in a variety of situations.

Experimentation on the Internet

Amazon, Google, and other dot-com companies have exploited the favorable properties of the web for relentless experimentation on the interaction between their users and websites. Google has set up a substantial infrastructure to run experiments quickly and cheaply on its search site.* For example, two search algorithms can be interleaved to compare results. The dependent (outcome) variables include the number of results clicked, how long the user continues to search, and other measures of

* Presentation by Hal Varian, Google chief economist, UCSD May 2007. See also Shankland (2008) and Varian (2006).

user satisfaction. Multiple page layout variables such as colors, white space, and the number of items per page are also tested. Through such tests, Google discovers the "real estate value" of different parts of the search page. Google tested 30 candidate logos for the Google checkout service, and, overall, it now runs up to several hundred experiments per day.*

Such experiments have excellent properties with respect to most of the criteria we have discussed. They have very fast information cycle times (a few days), very low variable cost because they use EVOP, and high fidelity because they are run in the real world. Even if individual improvements are very small, or more precisely, have low standard deviation, the S/N ratio can still be excellent because of very large sample sizes. Google runs experiments on approximately 1% of the relevant queries, which in some cases can give a sample of one million in a single day. This can reliably identify a performance improvement of only 1 part in 10,000.

However, Internet-based experiments still face potential problems on the focused/comprehensive dimension. In many experiments the goal is to affect the *long*-term behavior of customers; but in a single user session only their *immediate* behavior is visible. A change could have different effects in the short and long run, and sometimes even opposite effects. Thus, experiments that measure immediate behavioral effects are essentially short-loop experiments, with a corresponding loss of fidelity. A company like Amazon can instead track the long-term buying behavior of customers in response to minor changes, but such experiments are slower and noisier. It is also difficult to measure the long-term effects on customers who do not log in to sites.

Apple Breeding: Five-fold Reduction of Information Cycle Time

Breeding plants for better properties is an activity that is probably millennia old. Norman Borlaug's "Green Revolution" bred varieties of rice and other traditional crops that were faster growing, higher yielding, and more resistant to drought and disease. However, some plants, including apple trees and related fruits, have poor properties for breeding (Kean 2010). One difficulty is that apple trees take about five years to mature and begin to bear fruit, leading to very long experimental cycles. Finding out whether a seedling is a dwarf or full-sized tree can take 15 years. A second difficulty is that, because of Mendelian inheritance, traditional breeding produces a high percentage of undesirable genetic combinations, which cannot be identified until the trees begin bearing. Finally, the underlying genetic variation in American apples is small, as virtually all domestic plants are cloned from a small number of ancestors, which themselves originated from a narrow stock of seeds promulgated by Johnny Appleseed and others 200 years ago. As a result, apple-breeding experiments have been slow and expensive, with low variation in the underlying "ideas."

Botanists are now using four techniques to accelerate the rate of learning about apples. First, a USDA scientist collected 949 wild variants of the ancestor to domestic apples from Central Asia. This approximately doubled the stock of

* Erik Brynjolfsson, quoted in Hopkins (2010, 52). The number 50 to 200 experiments at once is given in Gomes (2008).

apple-specific genes available as raw material for learning (idea variety). Second, the time from planting to first apples is being shortened from five to about one year by inserting "fast flowering" genes from a poplar tree (information cycle time). Third, when two trees with individual good traits are crossed to create a combination of the two traits, DNA screening will be able to select the seedlings that combine both of the favored traits, without waiting until they mature. This reduces the number that must be grown by 75% (short-loop experiment to reduce cost). Finally, some of the wild trees were raised in a special greenhouse deliberately full of pathogens (specialized learning location, to improve cycle time and S/N).

Short-Loop Experiments for AIDS: The Critical Role of Measurement

The human immunodeficiency virus (HIV) is the virus that eventually leads to acquired immunodeficiency syndrome (AIDS), but the delay between the initial HIV infection and AIDS onset can be many years, even in untreated individuals. This delay made experiments on both prevention and AIDS treatment quite slow. It was not even possible to be sure someone was infected until they developed AIDS symptoms. Both controlled and natural experiments were very slow in consequence. In the early 1990s, new tests allowed a quantitative measurement of HIV viral loads in the bloodstream. Researchers hypothesized that viral load might be a proxy measurement for the occurrence and severity of infection, potentially permitting short-loop experiments that would be years faster than waiting for symptoms to develop. There is, however, always the question of the fidelity of short-loop experiments: is viral load truly a good proxy for severity of infection?

Using a controlled experiment to validate the viral load measure would be impossible for ethical and other reasons. As an alternative, a natural experiment consists of measuring viral loads in a number of individuals who may have been exposed, and then waiting to see what happens to them. Such an experiment would take years, and it would require a large sample to get a decent S/N ratio. Fortunately, it was possible to run a quick natural experiment using historical data (Mellors 1998). Mellors and colleagues measured viral loads in stored blood plasma samples from 1600 patients. These samples had been taken 10 years earlier, before the measurement technique existed. From the samples plus mortality data on the individuals, they calculated the survival and life expectancy rates as a function of the initial viral load. They found very strong relationships. For example, if the viral load was below 500, the six-year survival rate was over 99% versus 30% if the load was over 30,000.

This natural experiment on historical data achieved a high S/N ratio in a relatively short period. However, even with strong results, "correlation does not prove causation." In medicine the use of historical data is called a "retrospective trial." Fortunately, several large-scale prospective and controlled trials were able to demonstrate that if a treatment reduced viral load, it also reduced the likelihood of the disease progressing from infection to full AIDS. Therefore, viral load became a useful proxy for short-loop experiments on possible treatments, as well as for treating individual patients.

An ongoing example of developing a new measurement to permit shorter loop experiments is the recent discovery of seven biomarkers for kidney damage (Borrell 2010). These biomarkers permit faster detection and are sensitive to lower levels of damage in animal studies, and presumably for humans as well. This will allow for the better assessment of toxicity at an earlier stage of drug trials.

Both examples highlight how identifying new variables and developing practical measurements for them can accelerate learning rates several-fold. Especially good variables can predict final outcomes early, thus allowing for more focused experiments.

Fidelity Problems Due to Location in Clinical Trials

Sadly, searches for short-loop measurements that predict the course of illness are not always so successful. A particularly insidious example of the problem with short-loop experiments is the use of selective health effects as the outcome measurement in controlled clinical drug trials. If a drug is designed to help with disease X, in a large clinical trial should the outcome measure be symptoms related to X, clinical markers for X, mortality due to X, or measures of total health? Should immediate effects be measured, or should patients be followed over multiple years? The assumed causal chain is: new medication → clinical markers for disease X → symptoms for X → outcomes for X → total health of those taking the medication. The target outcome is the *total* health effect of the new medication, and if the disease is a chronic one that requires long-term treatment, this can take years to measure. Looking at just the markers for disease X, or even at the outcomes of disease X alone, are forms of short-loop experiments.

However, the problem with short-loop experiments is fidelity: Is the result for disease X a reliable predictor of overall health effects? In the human body, with its variety of interlocking homeostatic (feedback) systems, multiple complex effects are common. Even if X gets better, overall health may not. Although clinical trials are supposed to look for side effects, side effect searches generally have a low S/N. They are also subject to considerable conflict of interest, as the drug trials are usually paid for by the pharmaceutical company. There is ample evidence that this sometimes affects the interpretation and publication of study results (see, e.g., DeAngelis and Fontanarosa 2008).

A tragic example of this problem was the anti-diabetes drug, Avandia. It was one of the world's highest-selling drugs, with sales of $3.2 billion in 2006 (Harris 2010). Untreated diabetes raises the mean and variance of blood glucose levels, and better control of the glucose level is viewed as a good indicator of effectiveness in treating diabetes. However, once Avandia was put on the market the drug caused serious heart problems. In fact, one estimate was that it raised the odds of death from cardiovascular causes by 64% (Nissen and Wolski 2007). Given the already high risk of heart attacks for diabetics, this indicates that it decreased overall survival. Unfortunately, many drug trial reports do not even list overall mortality for the treatment and placebo populations, making it impossible to calculate the overall effects. Doctors are essentially forced to assume that the short-loop experiment has adequate fidelity for overall patient health.

Using intermediate disease markers for a focused clinical trial is not the only problem with the fidelity of clinical trials. Trial sample populations are carefully screened to improve the experimental properties for the primary (desired) effect. This is a move along the analytical/physical spectrum, because the trial population is not the same as the general population, which will take the drug if it is approved. For example, trials often reject patients who are suffering from multiple ailments or are already taking another medication for the targeted disease. For ethical reasons, they almost always avoid juvenile patients. These changes potentially reduce the fidelity of the experiment.

An example is the use of statins to reduce cholesterol and heart disease. While multiple clinical studies show that statins lower the risk of heart attack, they also have side effects that include muscle problems, impotence, and cognitive impairment (Golomb and Evans 2008). Few of the clinical trials of statins measure *overall* mortality, and the available evidence suggests that they improve it only in men with pre-existing heart disease. Most of the clinical trials have been conducted on this population, yet statins are now widely prescribed for women and for men with high cholesterol but without heart disease. The evidence suggests that for these populations, they do not actually improve overall mortality (Parker-Pope 2008). Since they also have side effects, this suggests that they are being widely over-prescribed.

Astronomy and other Observational Fields

In some situations, causal variables cannot be altered and so only natural experiments are possible. However, when observations are expensive and difficult, the learning process is very similar to a series of controlled experiments. Often the experiments require elaborate data collection, and the choice of where to observe and what to collect raises the same issues as setting up controlled experiments. The only difference is that there is usually no choice of location, which is at the "real world" end of the physical-analytical spectrum.*

The classic observational science is astronomy, since extra-planetary events cannot be altered, though the observation time on instruments is scarce and rationed, especially for wavelengths blocked by Earth's atmosphere. One approach to finding extra-solar planets is to look for periodic dimming in the light from a star. The magnitude of the dimming gives an indication of planetary size. The S/N ratio of the measurements is critical, as small planets will have little effect. Choosing stars to observe that are more likely to have planets and to have high S/N ratios is therefore crucial. Even so, a large sample is needed because if the planet's orbit is not on the same plane as our sun, no occlusion will be visible. A special satellite mission, "Kepler," has been launched for this purpose (Basri et al. 2005). A recent paper posits that the Kepler mission can also detect Earth-sized moons in habitable zones (Kipping et al. 2009). The authors used mathematical models of hypothetical moons to evaluate different detection strategies and their statistical efficacy.

* Even in observational fields like astronomy, controlled experiments are possible using mathematical models, though the results cannot be validated by controlled experiments in the real world.

CONCLUSIONS

This chapter shows how the rate of learning by experimentation—and, by extension, the slope of many learning curves—is heavily determined by the ways in which experiments are conducted. We identified five criteria that measure experimental effectiveness. They are: information cycle time (speed), S/N ratio, cost, value and variety of ideas, and fidelity. The statistical literature on experimentation deals formally with only one of these, the S/N ratio, but it offers insights that are helpful in dealing with the others. For example, the powerful method of *fractional factorial* experiments looks at the effects of multiple variables simultaneously rather than one at a time, thereby improving both speed and cost.

These five criteria are, in turn, largely determined by how the experiment is designed, and we discuss two high-level design decisions: location and type of experiment. "Location" has two orthogonal axes, referred to as analytical to physical, and focused to comprehensive. Generally, more analytical or more focused locations reduce the fidelity of the experiment, but improve other criteria such as information cycle time and cost. We discuss four types of experiments: controlled, natural, evolutionary, and ad hoc. Although controlled experiments are often viewed as inherently superior, for some types of problems they are impossible, and in other cases they are dominated by other types of experiments.

We have not discussed the meta-problem of designing good learning environments. For example, "just in time" manufacturing principles encourage just the key properties of fast information cycle time and good S/N ratio, while providing perfect fidelity (Bohn 1987). We have also said little about *what* is being learned, since it is heavily situation specific. However, G. Taguchi claimed that the knowledge being sought in experiments was often too narrow. He pointed out that in many situations the *variability* in outcomes is just as important as the mean level of the outcome. He proposed specific experimental methods for simultaneously measuring the effects of process changes on both mean and variation. For example, he suggested that pilot lines and prototypes should be built with material of normal quality, rather than using high-quality material to improve the S/N ratio of experiments. Taguchi's insight reminds us that "the first step in effective problem solving is choosing the right problem to solve."

ACKNOWLEDGMENTS

A multitude of colleagues, engineers, and managers have helped with these ideas over many years. Roger Bohn's research on this topic was supported by the Harvard Business School Division of Research, and the Alfred P. Sloan Foundation Industry Studies Program. His friends at HBS were especially influential, including Oscar Hauptman, Jai Jaikumar, Therese Flaherty, and Roy Shapiro. Gene Meiran and Rick Dehmel taught him about semiconductors. Jim Cook provided a number of comments and ideas. Michael Lapré's work was supported by the Dean's Fund for Faculty Research from the Owen School at Vanderbilt.

REFERENCES

Adler, P.S., and Clark, K.B., 1991. Behind the learning curve: A sketch of the learning process. *Management Science* 37(3): 267–281.

Applebaum, W., and Spears, R.F., 1950. Controlled experimentation in marketing research. *Journal of Marketing* 14(4): 505–517.

Basri, G., Borucki, W.J., and Koch, D., 2005. The Kepler mission: A wide-field transit search for terrestrial planets. *New Astronomy Reviews* 49(7–9): 478–485.

Bohn, R.E., 1987. *Learning by experimentation in manufacturing*. Harvard Business School Working Paper 88–001.

Bohn, R.E., 1994. Measuring and managing technological knowledge. *Sloan Management Review* 36(1): 61–73.

Bohn, R.E., 1995a. Noise and learning in semiconductor manufacturing. *Management Science* 41(1): 31–42.

Bohn, R.E., 1995b. The impact of process noise on VLSI process improvement. *IEEE Transactions on Semiconductor Manufacturing* 8(3): 228–238.

Bohn, R.E., 2005. From art to science in manufacturing: The evolution of technological knowledge. *Foundations and Trends in Technology, Information and Operations Management* 1(2): 129–212.

Borrell, B., 2010. Biomarkers for kidney damage should speed drug development. *Nature,* May 10. http://www.nature.com/news/2010/100510/full/news.2010.232.html?s=news_rss (accessed July 9, 2010).

Box, G.E.P., Hunter, J.S., and Hunter, W.G., 1978. *Statistics for experimenters*. New York: Wiley.

DeAngelis, C.D., and Fontanarosa, P.B., 2008. Impugning the integrity of medical science: The adverse effects of industry influence. *Journal of the American Medical Association* 299(15): 1833–1835.

Dutton, J.M., and Thomas, A., 1984. Treating progress functions as a managerial opportunity. *Academy of Management Review* 9(2): 235–247.

Golomb, B.A., and Evans, M.A., 2008. Statin adverse effects: A review of the literature and evidence for a mitochondrial mechanism. *American Journal of Cardiovascular Drugs* 8(6): 373–418.

Gomes, B., 2008. *Search experiments, large and small*. Official Google blog http://googleblog.blogspot.com/2008/08/search-experiments-large-and-small.html (accessed July 9, 2010).

Harris, G., 2010. Research ties diabetes drug to heart woes. *The New York Times*, February 19.

Hopkins, M.S., 2010. The four ways IT is revolutionizing innovation – Interview with Erik Brynjolfsson. *MIT Sloan Management Review* 51(3): 51–56.

Hunter, W.G., and Kittrell, S.R., 1966. Evolutionary operations: A review. *Technometrics* 8(3): 389–397.

Juran, J.M., and Gryna, F.M., 1988. *Juran's quality control handbook.* 4th ed. New York: McGraw-Hill.

Kean, S., 2010. Besting Johnny Appleseed. *Science* 328(5976): 301–303.

Kipping, D.M., Fossey, S.J., and Campanella, G., 2009. On the detectability of habitable exomoons with Kepler-class photometry. *Monthly Notices of the Royal Astronomical Society* 400(1): 398–405.

Lapré, M.A. 2011. Inside the learning curve: Opening the black box of the learning curve. In *Learning curves: Theory, models, and applications*. ed. M.Y. Jaber. Chapter 2. Boca Raton: Taylor & Francis.

Lapré, M.A., Mukherjee, A.S., and Van Wassenhove, L.N., 2000. Behind the learning curve: Linking learning activities to waste reduction. *Management Science* 46(5): 597–611.

Lapré, M.A., and Van Wassenhove, L.N., 2001. Creating and transferring knowledge for productivity improvement in factories. *Management Science* 47(10): 1311–1325.

Lee, H., Whang, S., Ahsan, K., Gordon, E., Faragalla, A., Jain, A., Mohsin, A., Guangyu, S., and Shi, G., 2003. *Harrah's Entertainment Inc.: Real-time CRM in a service supply chain*. Stanford Graduate School of Business Case Study GS-50.

Leonard-Barton, D., 1988. Implementation as mutual adaptation of technology and organization. *Research Policy* 17(5): 251–267.

Leonard-Barton, D., 1995. *Wellsprings of knowledge: Building and sustaining the sources of innovation*. Cambridge: Harvard Business School Press.

Little, J.D.C., 1966. A model of adaptive control of promotional spending. *Operations Research* 14(6): 1075–1097.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. *Nature* 453:869–878.

March, J.G., and Simon, H.A., 1958. *Organizations*. New York: Wiley.

Mellors, J.W., 1998. Viral-load tests provide valuable answers. *Scientific American* 279:90–93.

Murnane, R., and Nelson, R.R., 1984. Production and innovation when techniques are tacit: The case of education. *Journal of Economic Behavior and Organization* 5(3–4): 353–373.

Nelson, R.R., 2008. Bounded rationality, cognitive maps, and trial and error learning. *Journal of Economic Behavior and Organization* 67(1): 78–87.

Newell, A., and Simon, H., 1972. *Human problem solving*. Englewood Cliffs: Prentice-Hall.

Nissen, S.E., and Wolski, K., 2007. Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine* 356(24): 2457–2471.

Parker-Pope, T. 2008. Great drug, but does it prolong life? *The New York Times*, January 29.

Pearl, J., 2001. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Shankland, S., 2008. We're all guinea pigs in Google's search experiment. *CNET News*, May 29. http://news.cnet.com/8301-10784_3-9954972-7.html (accessed July 9, 2010).

Terwiesch, C., and Bohn, R.E., 2001. Learning and process improvement during production ramp-up. *International Journal of Production Economics* 70(1): 1–19.

Terwiesch, C., and Ulrich, K.T. 2009. *Innovation tournaments*. Boston: Harvard Business School Press.

Thomke, S.H., 1998. Managing experimentation in the design of new products. *Management Science* 44(6): 743–762.

Ulrich, K.T., and Eppinger, S.D., 2008. *Product design and development*. New York: McGraw Hill.

Varian, H.R., 2006. The economics of internet search. *Rivista di Politica Economica* 96(6): 9–23.

Weber, C., 2004. Yield learning and the sources of profitability in semiconductor manufacturing and process development. *IEEE Transactions on Semiconductor Manufacturing* 17(4): 590–596.