HARVARD | BUSINESS | SCHOOL

The Effects of Hierarchy on Learning and Performance in Business Experimentation

Sourobh Ghosh
Stefan Thomke
Hazjier Pourkhalkhali

Working Paper 20-081

The Effects of Hierarchy on Learning and Performance in Business Experimentation

Sourobh Ghosh
Harvard Business School

Stefan Thomke
Harvard Business School

Hazjier Pourkhalkhali
Optimizely

Working Paper 20-081

Funding for this research was provided in part by Harvard Business School.

The Effects of Hierarchy on Learning and Performance in Business Experimentation[i]

Sourobh Ghosh[ii], Stefan Thomke[iii]
Harvard Business School

Hazjier Pourkhalkhali
Optimizely

February 13, 2020

Abstract

Do senior managers help or hurt business experiments? Despite the widespread adoption of business experiments to guide strategic decision-making, we lack a scholarly understanding of what role senior managers play in firm experimentation. Using proprietary data of live business experiments from the widely-used A/B testing platform, Optimizely, this paper estimates the association of management hierarchy with learning from experiments and their performance outcomes across industries and contexts. Our findings suggest that senior management's association is mixed. On the one hand, senior managers' involvement associates with bolder experiments that create more statistically significant learning signals aiding in the exploration of new strategic directions. On the other hand, their involvement associates with less cause-and-effect learning that is instrumental to optimization and performance improvements. Our results contribute to a burgeoning literature on experimentation in strategy, while helping articulate limits that organizational design might place on data-driven decision-making. Furthermore, we describe different experimental learning modes in the formation of strategy, offering important implications for how managers can modulate search and performance outcomes.

Keywords: Experimentation; Innovation; Organization Design; Search; New Product Development

[i] We gratefully acknowledge feedback from Rory McDonald, Jan Rivkin, and Andy Wu. Special thanks to Charles Pensig and Optimizely for making data available for this research. Andrew Marder and Victoria Liublinska provided excellent research assistance. We acknowledge financial support from the HBS Division of Research and Faculty Development. All errors and omissions in this draft are those of the authors.

[ii] Address: Wyss House, 20 N Harvard St, Boston, MA 02163. Email: sourobh@hbs.edu

[iii] Address: Morgan Hall 489, 15 Harvard Way, Boston, MA 02163. Email: sthomke@hbs.edu

Introduction

Scholars of strategy, operations, organization, and entrepreneurship have noted the increasing importance of experimentation in strategy formation and innovation performance (Andries, Debackere, and Van Looy, 2013; Camuffo *et al.*, 2019; Gans, Stern, and Wu, 2019; Levinthal, 2017; McDonald and Eisenhardt, 2019; Thomke 2003). For example, the recent proliferation of A/B testing has made experimentation an attractive method by which startups can test various elements of their value creation and capture processes (Koning, Hasan, and Chatterji, 2019). The benefits of an experimental approach include faster learning, improved performance, and reduced errors in strategic decision-making (Camuffo *et al.*, 2019; Thomke, 1998). Thus organizations are scaling experiments to test strategic decisions, ranging from high tech and retail to non-profit organizations and even political campaigns (Thomke, 2020).

The proliferation of experimentation raises a tricky question about hierarchy: If major strategic decisions are made with the help of experiments, then what's the job of senior managers? On the one hand, they may help firms capture the learning and performance benefits of conducting business experiments. Senior managers are well-positioned within their organizations to drive the adoption of an experimental approach to searching for and validating business strategy (Levinthal, 2017; Thomke, 2020). They can authorize exploratory experiments that intentionally vary and select among strategic choices directing resources away from low-performing alternatives to high-performing ones (Burgelman, 1994). With the power to support experimentally-driven pivots (Camuffo *et al.*, 2019), executive attention may help create and capture learning and their performance benefits.

On the other hand, literature on judgement and decision-making suggests that senior management bias may hamper those learning and performance benefits. Senior managers may become overconfident in their beliefs as a result of their organizational power (Anderson and Galinsky, 2006; Fast *et al.*, 2009), preventing the exploration of new ideas. Furthermore, executives are less likely to revise their judgement in response to advice (Harvey and Fischer, 1997; See *et al.*, 2011), including those supported by test data. Indeed, practitioners have warned each other of executives whose opinions crowd out the role of experimental evidence in strategic decisions, deeming them the "Highest Paid Person's Opinions" (HiPPOs) (Kohavi, Henne, and Sommerfield, 2007). Given the contradictory influences of hierarchy, we ask: *how* does management seniority associate with learning and performance outcomes from experiments?

To address this question, we use a proprietary dataset of 6,375 business experiments run on the widely-used A/B testing platform, Optimizely. Our global dataset of live business experiments include start-ups, Fortune 500 companies, and global enterprises, and represents a wide range of industries and contexts, giving us reliable cross-sectional estimates of business experimentation practice. Furthermore, this paper is the first to use Optimizely's detailed microdata on experiments to generate measures of complexity and parallel testing in experimentation.

Our findings suggest that senior management's involvement in experiments is not as simple as the HiPPO warning or "more executive attention" advice. While all experimentation is about learning, we find that hierarchy is associated with different learning modes and performance outcomes. More senior management involvement is associated with bolder experiments, creating more statistically significant learning signals ("wins") that aid in the exploration of new strategic directions. But senior managers may also inadvertently undermine cause-and-effect learning that is instrumental to optimization and performance improvements.

This study offers a number of contributions to the strategy and organizations literatures. First, we contribute to a burgeoning literature on experimentation in strategy (Camuffo *et al.*, 2019; Levinthal, 2017) by linking hierarchy to performance outcomes in the study of experimentation. Second, our findings contribute to understanding the mixed effects that organizational design may place on data-driven decision-making in the digital age (Brynjolfsson and McElheran, 2016; Puranam, Alexy, and Reitzig, 2014). Despite the promise of large-scale experimentation, we find that hierarchy may insert bias in organizational search, favoring potentially lower-performing returns. Third, our empirical results describe different experimental learning modes in the formation of strategy, offering important implications for how managers can modulate firm search and performance outcomes.

Experimentation in Strategy Formulation

Firms have a long and storied history of using experiments to drive innovation, from Edison's famous Menlo Park laboratory (Millard, 1990) to 3M's culture of experimentation leading to the development of consumer products such as the Post-it Note (Nayak and Ketteringham, 1997). Since then, scholars have studied the role of experiments in industrial R&D, which has involved technologies such as prototyping, simulation, and combinatorial chemistry (Thomke and Kuemmerle, 2002). With the emergence of software-based testing and online customer interactions, business experimentation has entered a watershed moment. Inexpensive experimentation is no longer limited to R&D departments but now available to the entire firm and can be run in real-time, on live customers, to adjudicate firm-level commitments. This has profound implications for strategic management practice, as firms increasingly use experiments to test and form elements of strategy (Contigiani and Levinthal, 2019). Popularized by practitioner frameworks such as the "Lean Startup" (Ries, 2011), an experimental approach to entrepreneurial strategy has been adopted by organizations across contexts and vintage, ranging from financial services companies to hardware manufacturers. Mark Okerstrom, the former CEO of Expedia Group, underscored the strategic importance of running many experiments: "In an increasingly digital world, if you don't do large-scale experimentation, in the long term—and in many industries the short term—you're dead. At any one time we're running hundreds if not thousands of concurrent experiments, involving millions of visitors. Because of this, we don't have to guess what customers want; we have the ability to run the most massive 'customer surveys' that exist, again and again, to have them tell us what they want" (Thomke, 2020).

In contrast to other modes of strategy formation, experimentation is unique in that it balances the relative advantages and disadvantages of deliberate, cognition-driven strategy with emergent, action-driven strategy (Mintzberg, 1978; Ott, Eisenhardt, and Bingham, 2017). Emergent methods of strategy formation, such as bricolage and trial-and-error learning, prioritizes signals from the external environment, which is the ultimate arbiter of the fitness of a strategy. While these approaches may be effective at screening sets of choices activities for their performance, the firm is often left with an incomplete understanding of *why* a certain set of activities yields superior performance, as the firm itself is often not the ultimate source of variation among activities. In contrast, deliberate strategy prioritizes intentionality, where the firm fully controls its set of activities. Nonetheless, a fully deliberate approach absent some form of interim feedback from the external environment is often doomed to failure. Thus, an experimental approach to strategy formation offers a middle-ground between cognition and action (Levinthal, 2017), where the firm can balance cognition and its benefits of a holistic, causal understanding of strategy with the feedback and performance-screening benefits of action (Ott *et al.*, 2017).

Experimentation in strategy formulation helps the firm learn useful information about its available choices or activities (Gans *et al.*, 2019). For instance, entrepreneurs often find their strategy by learning through experiments rather than via traditional strategic planning (Carter, Gartner, and Reynolds, 1996; Murray and Tripsas, 2004; Ries, 2011). In particular, an experiment offers interim feedback on the fitness of a set of choices or activities (Levinthal, 2017). Under a classical planning approach, choices are validated via the long-term process of environmental selection; as a form of feedback, this information may arrive too late for a firm to act upon it. An experimental approach, which effectively screens for opportunities on the basis of interim feedback, helps ventures pivot faster and avoids choices that yield false-positive returns, or erroneous learning (Camuffo *et al.*, 2019). Experiments can also facilitate the identification of higher-performing choices (Gruber, MacMillan, and Thompson, 2008). In a study of A/B testing, a form of online experimentation, Koning et al. (2019) find that its adoption associates with increased product introductions and higher website performance over time.

The Effects of Hierarchy

To illustrate how management can influence the testing of strategic elements, consider the following example from the world's most visited travel accommodation platform Booking.com (Thomke and Beyersdorfer, 2018). In December 2017, just before the busy holiday travel season, the company's director of design proposed a radical experiment: testing an entirely new layout for the company's home page. Instead of offering lots of options for hotels, vacation rentals, and travel deals, as the existing home page did, the new one would feature just a small window asking where the customer was going, the dates, and how many people would be in the party, and present three simple options: "accommodations," "flights," and "rental cars." All the content and design elements—pictures, text, buttons, and messages—that Booking.com had spent years optimizing would be eliminated. Booking.com runs more than 1,000 rigorous tests simultaneously and, by one estimate, more than 25,000 tests per year (Thomke, 2020). At any given time, quadrillions (millions of billions) of landing-page permutations are live, meaning two customers in the same location are unlikely to see the same version. What could Booking.com possibly learn from the experiment proposed by this senior manager? Too many variables would have to change at the same time, making cause-and-effect learning about individual design elements virtually impossible. Instead, the design director positioned the experiment as "exploratory"—testing the significance of a new landing page design and elements of a new business strategy that resembled an emerging competitor: Google. And, as we will see in this paper's results, it is less likely that such an experiment would have been proposed by a team that consists of only junior employees. The example of Booking.com's landing page experiment illustrates the influence that senior managers can exert on a firm's learning strategy. But the impact of their influence is ambiguous: scholars have found that hierarchy can benefit and impede decision-making.

Benefits of Hierarchy

Executive attention has been shown to be a powerful influence in motivating exploratory change in many contexts (Gavetti *et al.*, 2012; Ocasio, 1997), such as when introducing technological innovations (Kaplan, 2008), increasing technological responsiveness to competitors (Eggers and Kaplan, 2009), and the adoption of expansive global strategies (Levy, 2005). In the context of experimentation in organizations, lower-status individuals in the organization reduce their fear of failure and are found to experiment more often when senior managers clearly articulate values and incentives favoring experimentation (Lee *et al.*, 2004). At its core, experimentation allows managers to address the resource allocation problem in strategy

(Gavetti *et al.*, 2012). The problem can be framed as a tension between exploration and exploitation, where firms generally trade-off the search for new strategies against profiting from more certain returns with existing strategy (March, 1991). Rather than limiting themselves to a regime of pure exploration or exploitation, experimentation enables managers to attain a more efficient allocation of resources by splitting the difference in practice—enabling firms to explore new, potentially profitable strategies while remaining committed to an ongoing course of action, such as the optimization of such commitments. This idea of remaining "engaged in one course of action but cognizant of other possibilities" is embodied by the "Mendelian" executive, whose role is to foster an experimental approach to strategy (Levinthal, 2017).

The idea of the "Mendelian" executive is to draw organizational attention (Ocasio, 1997) to experimentation and the controlled exploration of novel, potentially high-performing strategies. Thomke (2020) finds that an organizational culture of experimentation begins with senior management awareness of the superiority of an experimental approach. When senior managers direct their attention towards and support an experimental approach to strategy, the firm may benefit from different learning modes and performance outcomes. For instance, an experimental approach to strategy is shown to reduce false positive learning mistakes while stimulating strategic pivots towards higher performance (Camuffo *et al.*, 2019). In particular, hierarchy may potentially induce improved selection of ideas (Knudsen and Levinthal, 2007), by offering additional checks on ideas and proposals that may prove to be maladaptive or lower performing (Keum and See, 2017). Furthermore, to avoid organizational myopia, a Mendelian executive may look to support decision-making structures that tolerate a variety of beliefs to promote the generation and selection of the highest performing alternatives available to the firm (Levinthal, 2017). From their position of influence, senior managers have the power within organizations to gradually direct the resource allocation mix towards higher-performing strategies (Burgelman, 1994), via support for a process of the reasoned generation and selection of alternatives from experimentation (Levinthal, 2017).

Impediments of Hierarchy

In spite of these potential benefits of involving executives in experimentation, commonly cited risks of senior involvement include introducing decision-making biases into the experimental process, such as overconfidence in prior beliefs. For instance, executives may suffer from biases due to their power within organizations, which has been shown to increase confidence, optimism, and a sense of control over future events (Anderson and Galinsky, 2006; Fast *et al.*, 2009). In turn, these power-driven biases may lead executives to rely too heavily on their own beliefs rather than considering experimentation to update their beliefs. Furthermore, research on advice taking suggests that decision makers, especially those in positions of power or authority, are less likely to revise their initial judgment in response to advice from others, leading to poor decisions (Harvey and Fischer, 1997; See *et al.*, 2011). A senior manager sitting atop an organizational hierarchy may be more likely to apply selection criteria which reflects the past but is maladaptive in the current moment (Aldrich, 1979). Together, these influences suggest that senior managers may not readily support experiments with the potential to challenge their current beliefs.

In communities of experimentation practice, the phenomenon of senior management biases impeding decision-making is given the moniker: the Highest Paid Person's Opinion (HiPPO) effect (Kohavi *et al.*, 2007). HiPPOs are associated with executives who, through their position of influence, advance potentially poor decision outcomes. Jim Barksdale, former CEO of Netscape, reportedly quipped his decision-making heuristic as “If we have data, let’s look at data. If all we have are opinions, let’s go with mine.”

When senior managers have overconfidence in their beliefs, the firm may suffer from impaired learning and performance. In studying how information is passed up to management by subordinates within firms, Reitzig and Maciejovsky (2013) cite subordinates’ fear of a lack of control over final outcomes and their apprehension of being negatively evaluated by superiors as reasons for reduced information sharing from subordinates to senior managers. Thus, information that could help the firm may not be sent up the hierarchy if it challenges a manager’s personal viewpoint. Similarly, Knudsen and Levinthal (2007) note that firms with an accurate screening ability of alternatives, such as those that adopt precise data-driven experimentation, should complement this capability with a managerial structure of polyarchy rather than that of hierarchy, which has a tendency to prematurely stop search. This may trap the firm in local performance peaks, harming firm performance (Levinthal, 1997). Csaszar (2012) finds empirical evidence supporting this view, where hierarchy in financial firms leads to errors of omission (foregoing investment in profitable projects) and fewer approved projects overall.

Methods

Our study addresses the effects of hierarchy on A/B testing, a form of experimentation. In an A/B test the experimenter sets up two experiences: “A,” the control, is usually the current system and “B,” the treatment variant, is some modification that attempts to improve something. Users are randomly assigned to the experiences, and key metrics are computed and compared. (*A/B/n tests* and *multivariate tests,* in contrast, assess more than one treatment variant or modifications of different variables at the same time.) Online, the modification could be a new feature, a change to the user interface (such as a new layout), a back-end change (such as an improvement to an algorithm that, say, recommends books at Amazon), or a different business model (such as free services, pricing models, or entirely new products and services) (Kohavi and Thomke, 2017). Even incremental variants can be effective at screening for high-performance innovations (Levinthal, 2017), especially in light of fat-tailed distributions which may yield large payoffs (Azevedo *et al.*, 2019).[1]

A/B/n testing is a particularly useful setting to study the role of hierarchy in strategic experimentation for several reasons. First, as seen in the Booking.com example, senior managers get involved in A/B/n testing, from minor improvements to entire website redesigns, because of their importance to online commerce (Thomke 2020). Second, an "experiment" is clearly defined—at least one treatment variant is tested against a control—separating it from other methods used by strategy and entrepreneurship practitioners such as effectuation and trial-and-error learning (Camuffo *et al.*, 2019; Ott *et al.*, 2017). The use of controls, combined with randomization, is particularly effective for cause-and-effect learning (Rosenbaum, 2017). Third, an experiment's design and search space choices are fully transparent in A/B/n testing, helping us assess the impact of a set of design choices that senior managers may potentially act through in order to influence learning and performance outcomes.

Data

We obtained access to a proprietary dataset from the third-party A/B/n testing platform, Optimizely, which supports more than one thousand clients across industries (e.g., retail, media,

[1] The opportunity of fat-tailed distribution was noted by Jeff Bezos, CEO of Amazon, in his 2015 letter to shareholders: "Outsized returns often come from betting against conventional wisdom, and conventional wisdom is usually right. Given a ten percent chance of a 100 times payoff, you should take that bet every time. But you're still going to be wrong nine times out of ten. We all know that if you swing for the fences, you're going to strike out a lot, but you're also going to hit some home runs. The difference between baseball and business, however, is that baseball has a truncated outcome distribution. When you swing, no matter how well you connect with the ball, the most runs you can get is four. In business, every once in a while, when you step up to the plate, you can score 1,000 runs." (SEC Archives, 2016: Letter to Amazon Shareholders from CEO Bezos.)

technology, travel, finance, etc.). While companies such as Google, Amazon, and Booking.com built in-house platforms years ago, Optimizely helped pioneer easy-to-use A/B/n testing tools for technical and non-technical business professionals. As a result, data from Optimizely provides access to experimentation practice across industries, contexts, organizational scale, and levels of technological sophistication.

Optimizely archives data from A/B/n experiments run on its cloud platform, including *p*-values, effect sizes, number of website visitors within an experiment, and experiment duration. Furthermore, the company collects detailed job role and rank data on users, enabling us to construct measures of organizational hierarchy within experimentation accounts registered with Optimizely. In addition, the company offers seven types of web element changes (e.g., HTML code changes, inserting an image, etc.) and records these changes as they are made.

Our unit of analysis is the experiment: an A/B/n test. Each test is an opportunity for the firm to learn and improve business performance, such as higher rates of customers who purchase products on a website. Not all tests on Optimizely's platform qualify as true experiments. For instance, the Optimizely interface enables "hotfix" behavior, which includes software patches for bugs in production software or rapid deployments of feature, content, or design changes that bypass formal release channels.[2] A basic requirement for an experiment is that it generates information that helps the firm learn (Thomke, 2003: pg. 98) and that is relevant to a firm's strategy or configuration of choices.[3]

[2] https://community.optimizely.com/t5/Testing-Ideas-Strategy/Ways-to-use-Optimizely-outside-of-just-A-B-testing/td-p/9406

[3] Rivkin (2000) notes that "a strategy is realized in the marketplace as a set of choices that influence organizational performance." For online platforms, it is the set of choices that affect how they interact and do business with their customers through a company's landing page.

To qualify as a true experiment for our analysis, an A/B/n test must meet the following criteria: 1) at least one change per treatment variant (i.e., no A/A tests)[4], 2) at least one treatment variant per test (i.e., not hotfixes), and 3) at least 1,000 website visitors per week that are allocated across control and at least one treatment variant (i.e., a meaningful sample size to power experiments). Therefore, an experiment ends during the last observed week in which traffic surpasses 1,000 visitors. Furthermore, we define outcomes at the experiment level (e.g., statistical significance, lift, etc.) for the primary metric, which an organization selected as its most important performance indicator on Optimizely's platform. Applying these criteria to the entire dataset yields a sample of 6,375 experiments run April 2018 to November 2018 from Optimizely for our analysis.[5]

Measures

Dependent Variables: Max Lift and Positive Statsig

The first dependent variable measures "lift," which is the *net improvement* that results from an experimental treatment. In particular, lift measures the percent improvement in the conversion rate for a key performance indicator of interest and is widely used in business experimentation practice (Gordon *et al.*, 2019). An example for an e-commerce website may be the percentage of users who complete a purchase of all the users landing on the shopping cart page, thus converting website visitors into paying customers. Thus, lift often has a direct impact on firm performance. We measure *Max Lift*, which represents the maximum lift on the primary metric across *n* variants of an experiment, as this represents the option or variant that is

[4] In an A/A test the current practice is compared with itself. A/A tests are used to check the quality of an experimentation infrastructure. If the *p*-value (false positive) is set to 0.1, one out of ten A/A tests should result in statistical significance.

[5] The sample window of experiments was chosen by Optimizely's data warehouse team. They considered the completeness, availability and quality of its raw data but did not analyze any of it.

highlighted on Optimizely's testing platform and most likely to be implemented after an A/B/n test.[6]

The second dependent variable, *Positive Statsig*, is any positive, statistically significant lift on the primary metric in an experiment: a *signal* that the observed treatment effect is unlikely the result of chance.[7] Framed as a "win" by A/B/n testing practitioners, a *Positive Statsig* result is a key performance indicator for the success of an individual experiment. An Optimizely account user would recognize a win graphically (green color). For experimenters, *Positive Statsig* signals that the treatment was worth exploring and builds confidence in a positive ROI from further rounds of experiments. Crossing the significance threshold also reduces the chance of making a false-positive error when evaluating strategic pivots (Camuffo *et al.*, 2019). This promotes quality in organizational learning, which is a change in the organization's knowledge or beliefs as a function of experience (Argote and Miron-Spektor, 2011; Puranam *et al.*, 2015). In an A/B/n test, the organization's existing knowledge is coded into the baseline or control variant. When an experiment yields a *Positive Statsig* signal, firms receive positive, statistically significant evidence to help update prior beliefs.

[6] An alternate measure of lift is the maximum statistically significant lift. We choose to operationalize lift in terms of its raw number rather than lift conditioned on statistical significance for three reasons. First, raw lift is a meaningful, interpretable outcome that is often used in practice to drive decision to implement changes, regardless of whether or not such a lift was deemed statistically significant. Second, conditioning lift on significance would correlate with our second dependent variable: *Positive Statsig*. We want to analyze learning modes and performance outcomes that are independent from another. Third, conditioning lift on significance may underestimate negative potential impacts of experiment treatments, such as those that lead to losses in conversion rate. While many negative effects due to experimental treatment are not strong enough to be deemed statistically significant, a negative lift represents real losses in conversion for firms. To adequately capture this risk, it is important to construct a measure of lift that does not censor out these potential losses.

[7] The treatment effect is the difference between the two sample averages (A and B). Given the multiple comparisons problem of multivariate A/B/n testing (where multiple treatment variants are compared to a control variant), Optimizely does not declare significance by calculating unadjusted *p*-values and comparing them to standard significance thresholds, as this would exacerbate the chance of making Type I errors. Instead, Optimizely employs false discovery rate (FDR) control using the Benjamini-Hochberg procedure (see Pekelis *et al.*, (2015) for further details). Significance is therefore reported to Optimizely users if 1 – FDR > 90%. Here, the use of 90% reflects standard industry practice of using a 10% threshold to deem statistical significance.

Independent Variable: Max Seniority

The involvement of senior managers in an experiment is measured by the variable *Max Seniority*. Hierarchy, brought about by the assignment of formal authority in organizations (Bunderson *et al.*, 2016; Keum and See, 2017), has been measured in a variety of ways, including span of control (Rajan and Wulf, 2006; Reitzig and Maciejovsky, 2013), tallness (Dalton, D.R., Todor, W.D., Spendolin, M.J., Fielding, G.J. and Potor, 1980; Hall and Tolbert, 2005; Lee, 2020) and centralization (Hage, 1965; Scott, 1998; Tannenbaum *et al.*, 1974). Our research captures the effect of increasing steepness of a hierarchy, which comes from the larger asymmetries in members' power, status, and influence (Anderson and Brown, 2010). To capture this effect, *Max Seniority* measures the highest rank of all individuals associated with an Optimizely project team. Users on an experimentation project team select their roles according to six standardized hierarchical levels, ranging from ``Specialist/Associate'' (ranked as a minimum value of 1) to ``C-Level/President'' (ranked as a maximum value of 6). Thus, an experiment associated with five Specialist/Associates would be coded as having Max Seniority of 1, whereas an experiment with three Specialist/Associates, a Vice President, and a CEO would be classified as having a *Max Seniority* of 6.[8]

Control Variables

We control for *Traffic* and *Duration*, which represent the number of web visitors included in an experiment (in thousands), and the number of weeks an experiment has been run, respectively. Both variables relate to the experiment's power and may influence the ability to

[8] The full standardized ranking is as follows: 1) Specialist/Associate, 2) Developer, 3) Coordinator, 4) Manager, 5) Vice President/Director, 6) C-Level/President. In the two provided examples, the team with *Max Seniority* of 6 has greater steepness than the team with Max Seniority of 1, in which there are no asymmetries in power, status, and influence from job roles. Given the prevalence of Specialist/Associates across all Optimizely user accounts, a higher *Max Seniority* score captures greater steepness between the highest-ranking individual and the lowest-ranking individuals, the Specialist/Associates on that given Optimizely user account.

detect statistical significance. In addition, we add week fixed effects to control for seasonal factors that might influence experimental outcomes.

At the firm level, we control for *Firm Age* through years since founding and *Firm Size* through the number of employees, both of which have been associated with firm search and innovation capabilities (Damanpour and Aravind, 2012). We also control for *Technological Integrations*, which measures the number of integrated technologies that Optimizely has detected when clients use its A/B/n testing platform (e.g., plug-ins to aid data analytics). This helps control for the technological sophistication of the firm, which may influence the value they derive from A/B/n testing. Finally, we include fixed effects to control for industry-driven heterogeneity across experiment outcomes. Descriptive statistics and pairwise correlations are shown in Table 1.

Insert Table 1 about here.

Model Specification

Our analysis of hierarchy's association with experimental outcomes is done using models of the following specification:

$$Y_i = \beta(Max\ Seniority_i) + X_i B + \eta_i + \delta_i + \epsilon_i$$

where Y_i is a performance measure of interest for experiment *i* (e.g., Max Lift), $Max\ Seniority_i$ is the most senior rank of individual associated with experiment *i,* X_i is a vector of controls associated with an experiment, and η_i and δ_i represent fixed effects for the industry and final week associated with experiment *i.* Our coefficient of interest is β, which estimates the association of maximum hierarchical rank and experimental outcomes. We estimate models using ordinary least squares with robust standard errors clustered to the team level.[9]

Results: The Effects of Hierarchy

Table 2 reports associations between increasing management seniority and outcomes. Model 2-1 shows that an increase in the hierarchical rank of the most senior person is associated with a 0.9% decrease (p = .016) in the conversion rate of an experiment.[10] However, Model 2-2 also shows that each increase of the hierarchical rank of the most senior person on an experimentation team is associated with a 1% increase (p = .047) in the chance of finding a positive, statistically significant learning signal (a "win").

Insert Table 2 about here.

The results from Table 2 present a paradox: that is, we would generally expect that higher lift correlates with higher rates of statistically significant outcomes. This intuition follows from an understanding of statistical power—where studying larger effect sizes would yield an improved chance of detecting significant effects. Nonetheless, it is possible that hierarchy's countervailing associations with lift and positive statsig may be the result of other mechanisms, which we explore in the following section.

[9] For the binary variable of *Positive Statsig*, we estimate models according to ordinary least squares rather than logit or probit for ease of interpretation and to avoid a potentially inconsistent maximum likelihood estimator in the presence of fixed effects (Greene, 2004). Nonetheless, our results for *Positive Statsig* are robust to the use of logit or probit models.

[10] Note that $lift = \frac{Treatment-Baseline}{Baseline}$. When performing a natural logarithm transform, we have ln $(lift + 1)$ = $ln\left(\frac{Treatment-Baseline}{Baseline} + 1\right)$ = $ln\left(\frac{Treatment}{Baseline}\right)$. Thus, the interpretation of our coefficient is a percent change in the baseline conversion rate.

Hierarchy and Design Choices

Design choices in experimentation can have a meaningful impact on learning modes and performance outcomes (Loch, Terwiesch, and Thomke, 2001; Sommer and Loch, 2004; Thomke, von Hippel and Franke, 1998). Is the mixed effect of hierarchy on performance outcomes above related to design choices and does senior management exert influence on learning modes through these choices? To find out, we examine two important design choices in A/B/n tests: the number of simultaneous changes in a treatment variant (which relates to the complexity of an experiment) and the number of variants that are tested in parallel.

Complexity of Treatment Variants

In strategy and organizational theory, complexity arises from several choices whose contribution to performance depend on one another (Levinthal, 1997; Simon, 1962). Thus a strategy can be thought of as a complete configuration of interdependent choices (Rivkin, 2000) and testing a new strategy requires multiple, interdependent changes to be made simultaneously (Pich, Loch, and Meyer, 2000, Rivkin 2000). More complex experiments can also signal the strength of a strategic direction as they explore new value landscapes (discover new "hills" of strategic value) and avoid getting stuck in local optimization (climbing existing "hills").

The downside of complex tests is that they are harder to interpret since many simultaneous changes make cause-and-effect learning problematic (Thomke, 2003). To facilitate ease of interpretation and to understand cause-and-effect relationships, approaches to entrepreneurial experimentation often prescribe testing one change at a time (Camuffo *et al.*, 2019), a heuristic that supports learning via incremental changes. Moreover, testing complex, tightly-coupled ideas may generate performance failures for the firm (Levinthal, 1997). For instance, Gavetti and Levinthal (2000) show that large cognitive realignments, which represent complex, interdependent changes in the space of action, can lead to immediate, short-term performance losses. In summary, increasing the complexity of treatment variants can encourage an exploratory, discovery-driven learning mode of strategic choices but inhibit cause-and-effect learning that is needed for the incremental optimization of commitments.

Number of Treatment Variants

Firms must also decide how many treatment variants (or options) are tested in parallel. Parallel testing arises when multiple treatments are tested simultaneously against a control. Having access to more variants may facilitate teams to consider alternatives that otherwise would be dismissed. The decreasing economic cost of testing (Koning *et al.*, 2019, Thomke 1998) favors such parallel testing, which is associated with higher short-run performance (Loch *et al.*, 2001). For instance, Azevedo *et al.* (2019) show that under fat-tailed distributions common in A/B testing experimentation, a lean approach of more tested interventions is preferred to help screen for extreme performance gains. Furthermore, a parallel testing approach enables changes to be allocated across multiple variants, which facilitates cause-and-effect learning. With fewer changes allocated to each variant, an experimenter can discern the source of an effect with greater efficiency.

Despite these near-term performance benefits, parallel testing may reduce power in testing while leading to the costly, excessive exploration of new alternatives. For a fixed sample size, an experiment with more variants tested against a control will feature a smaller sample for each variant.[11] This reduces power in testing, decreasing the chance that the real effect of a treatment is detected (true positive). In addition, testing multiple treatments in parallel increases

[11] Note this assumes that the experimenter is not leveraging variants to create a factorial experimental design, where sample size for a factor of interest is distributed across multiple variants (Czitrom, 1999; Montgomery, 2013). In interviews with Optimizely, we find that most A/B testers do not take a factorial design approach to their experiments.

the chance of inference error due to increasing false discovery rates that arise in multiple hypothesis testing[12] (Pekelis, Walsh, and Johari, 2015). Besides the threats of reduced power, parallel testing may also impede strategic commitment by prompting firms to excessively explore their alternatives (Gans *et al.*, 2019). For instance, Billinger *et al.* 2014 demonstrate in a laboratory study that human decision-makers exhibit a tendency to excessively explore new alternatives when they could focus on improving existing alternatives. Here, it is possible that parallel testing could distract organizations from improving existing products that could unlock higher performance potential.

Other Independent Variables: Variant Complexity and Variant Count

To examine how hierarchy associates with design choices, we used the Optimizely dataset to construct measures of the complexity and number of treatment variants. First, we measure variant complexity, or the total number of distinct change classes activated within an experiment. Each treatment variant tests a change from the baseline control, which is recorded as seven interdependent change classes on Optimizely's experimentation platform.[13] The complexity of the change tested increases with the number of interdependent change classes activated. For instance, a simple change to background color would count as one change, whereas a new checkout page could be composed of four distinct change classes. We construct two related measures of variant complexity—*Max Variant Complexity*, which measures the number of distinct change classes activated in the most complex variant within an experiment, and *Mean Variant Complexity*, which is the average variant complexity across all variants within an experiment.

[12] Here, each treatment arm would be testing a different alterative hypothesis.

[13] These change types are: 1) HTML code change, 2) HTML attribute changes, 3) Custom CSS code change, 4) Custom code change, 5) Insert/edit widgets, 6) Insert/edit images, 7) URL redirect changes.

To observe parallel testing, we measure *Variant Count*, or the number of treatment variants that are run in parallel within an individual experiment. A variant is a treatment to be tested against a control and a simple A/B test would be coded as two variants. Companies can run *n* simultaneous treatments to test of sets of interdependent choices that may define a strategy (Rivkin, 2000).

Results: Hierarchy and Design Choices

To understand the mixed effect of hierarchy on performance outcomes, we examine associations of increasing seniority with design choices in Table 3. Model 3-1 tests the association between hierarchy and *Max Variant Complexity*. In particular, we find that each increase of the hierarchical rank of the most senior person on an experimentation team is associated with 0.017 (*p* = .036) more distinct change classes in a treatment variant. In terms of parallel testing, we find in Model 3-2 that an increase in max seniority is associated with 0.037 (*p* = .038) fewer treatment variants tested per experiment. Given that variant complexity may also be achieved with more treatment variants, in Model 3-3, we test the association between hierarchy and *Mean Variant Complexity*, which measures the average number of simultaneously tested elements across variants in an experiment. Similar to Model 3-1, we find that an increase in senior rank associates with an increase variant complexity, or 0.018 (*p* = .018) more mean changes per experiment.

Insert Table 3 about here.

To understand how the aforementioned design choices may influence learning modes and performance outcomes, we turn to Table 4. Model 4-1 shows that an increase in the number of treatment variants within an experiment associates with a 3.8% increase in the conversion rate of an experiment (*p* < .001). This result confirms basic intuition that more treatment variants increase the chance of finding higher lifts.[14] Model 4-2 demonstrates the association between max variant complexity and lift, demonstrating a positive but smaller association with lift (i.e., a 1.4% in the conversion rate of an experiment, $p = 0.079$). Taken together, we find that although both variants count and variant complexity have positive associations with lift, when comparing effect sizes and p-values, our analysis suggests that *Variant Count* has a stronger, more noteworthy positive association with maximum lift.

Models 4-3 to 4-5 illustrate the association between design choices and positive statsig signals. Model 4-3 demonstrates that increasing the number of treatment variants has an association with *Positive Statsig* that is difficult to distinguish from zero ($p = .198$). Models 4-4 and 4-5 test the association between experimental complexity and the chance of a positive detection. These models show that a one-unit increase in *Max Variant Complexity* and *Mean Variant Complexity* associate with a 1.9% ($p = 0.041$) and 2.1% increase ($p = 0.032$) in the chance of a *Positive Statsig* learning signal, respectively.

Insert Table 4 about here.

Taken together, these results demonstrate the senior management involvement has a mixed effect on learning modes and performance. Increased hierarchy in experimentation teams is associated with increased variant complexity and positive, statistically significant performance signals found in discovery-driven learning. However, in favoring statistically significant signals, hierarchy may inhibit incremental, parallel cause-and-effect learning.

[14] To adjust for the multiple comparisons problem and increasing chance of committing Type I errors, Optimizely applies false discovery rate (FDR) control via the Benjamini-Hochberg procedure in the calculation of its results. Further detail on the calculation can be found here (Pekelis *et al.*, 2015). While our calculation of raw lift is not conditioned on statistical significance, it is important to note that given FDR control, there is little incentive to A/B/n testing practitioners to test more variants in the hope of artificially finding significant results.

Robustness Checks

We conducted several analyses to probe the study's robustness. First, we examine whether the results are sensitive to the chosen unit of analysis: the experiment. In experimentation programs, learning modes and performance outcomes are defined at the level of an individual A/B/n test. For each experiment, firms must decide whether to include senior managers, whose time and attention are limited. Thus, in practice management seniority is assigned at the level of an individual experiment. But could experiments associated with the same organization be dependent observations?

To address this question, we aggregate our analyses of hierarchy at the level of the experimentation team and show the results in Table A1. Model A1-1 demonstrates that increasing seniority is associated a 4.6% decrease ($p = 0.039$) in conversion rates*,* while Model A1-2 shows a 4.7% increase ($p < 0.001$) in the chance of *Positive Statsig* across experiments. Regarding design choices, Models A1-3 and A1-4 demonstrate that increasing seniority is associated with 0.077 ($p < 0.001$) more distinct change classes across experiments and 0.066 ($p < 0.001$) fewer variants on average across experiments. These findings demonstrate the robustness of our prior findings, with stronger and larger effect sizes when aggregated at the team level. While aggregating analyses at the organizational level sacrifices granularity of control at the level of the individual experiment (such as exact traffic, duration, week, etc.), results are also less prone to variability in results across individual tests (such as the quality of the individual idea being tested) which may influence outcomes such as lift and statistical significance. Estimates of effects at the team level help absorb some of this variability.

An alternative explanation may be that greater A/B/n testing tenure is driving our findings—insofar as more senior managers become involved with testing as the organization gains experience with experimentation. In Table A2, we control for *Experimental Experience*, which is each organization's total tenure on the Optimizely platform measured in number of days prior to a focal experiment. Associations between *Experimental Experience* and outcomes of learning and performance are difficult to distinguish from zero. Furthermore, we do not see any declines in the strength of association between max seniority and outcomes of *Max Lift* and *Positive Statsig*.

Finally, we examine the extent to which our findings are potentially influenced by diminishing marginal returns—and whether this property of diminishing returns may associate with hierarchy's relationships with lift and positive statsig learning signals. A/B/n testing practitioners have made the observation that effect sizes of experimental treatments can decrease over time.[15] This could be due to a variety of reasons, such as the initial novelty of the treatment wearing off on customers (Dmitriev *et al.*, 2017), the maturity of testing initiatives, general equilibrium effects (Heckman, Lochner, and Taber, 1999), or even the fact that websites become increasingly optimized over time and yield fewer opportunities for improvement. In Table A3, we re-run our analyses from Table 2 but control for the number of A/B/n tests run prior to a focal experiment with the measure *Number Prior Experiments*. Regarding lift, we find that after controlling for prior experiments, the effect size of increasing seniority in hierarchy on lift decreases somewhat (from -0.9% in Model 2-1 to -0.7% in Model A3-1), although the relationship between seniority and decreased lift retains a similar level of statistical significance ($p = 0.046$). Nonetheless, we find that controlling for the number of prior experiments has little discernible influence on the association between seniority and a positive statsig learning signal. Together, these results demonstrate the robustness of our main analyses in Table 2,

[15] https://www.semrush.com/blog/ab-testing-landing-page-elements/

demonstrating hierarchy's positive association with positive statsig learning outcomes and negative association with lift.

Discussion

We have found a mixed effect in hierarchy's association with experimentation: whereas senior managers associate with discovery-driven learning and statistically significant learning signals, they negatively associate with cause-and-effect learning and lift. We also found that senior managers' influence may flow through experimental design choices: more complex and fewer treatment variants. So what are some possible explanations for the differences between experiments with senior and junior participation (i.e., experiments that lack a senior manager)? And why might senior managers associate with one set of outcomes versus another?

Here, it may be useful to think of each experiment as searching a rugged value landscape topography, where "hills" represent choices and their payoffs (Levinthal, 1997; Thomke, von Hippel, and Franke, 1998). In this search, seniority in management decides *which* hill to climb versus gradually climbing an existing hill that an organization has settled on. By encouraging experiments with complex, high-degree changes, senior managers can guide longer jumps in the landscape (Levinthal, 1997). In contrast, experiments with senior management participation are more likely testing incremental changes in parallel. On average, senior-driven experiments may "win" more by detecting significant effects, giving confidence to a manager who wishes to avoid false-positive returns (Camuffo *et al.*, 2019) and ensure that they are climbing an appropriate hill. However, the risk of committing large, simultaneous changes (i.e., a long jump) is performance losses when compared to more incremental hill-climbing (Gavetti and Levinthal, 2000). It may seem counterintuitive that smaller scope changes lead to higher average performance, but this idea is not new. In fact, incremental moves may help unlock performance discontinuities and success. As Levinthal (2017) points out, "…many instances of dramatic strategic change or success can be understood at a fine level of granularity as being relatively incremental in the space of action…seemingly rapid technological change is the consequence of fairly incremental moves in technological space, with the seemingly discontinuous change stemming from a shift of the technology to a new niche or application domain." In a recent study of United States economic growth, the authors estimated that, between 2003 and 2013, improvements to already existing products accounted for about 77 percent of increases in growth (Garcia-Macia, Hsieh, and Klenow, 2019). Similarly, studies in manufacturing and computer technology have shown that significant performance advances were often the result of minor innovations (Hollander, 1965; Knight, 1963).

While more complex experiments associated with senior managers may result in lower average returns in lift, it is also possible that senior managers may be interested in long-term performance outcomes not captured by the data in the present study. This would align with the notion of senior managers screening the broader landscape of opportunities for which hill to climb, with the promise of greater returns in the long-run. In contrast, junior-driven teams may have near-term incentives in experimentation, influencing their search behavior (Lee and Meyer-Doyle, 2017). Measuring the effect of hierarchy and design choices on long-term performance outcomes is an important area for future study.

Conclusion

The proliferation of business experimentation in strategy formation has emerged as an important area of study (Azevedo *et al.*, 2019, Camuffo *et al.*, 2019; Koning *et al.*, 2019; Levinthal, 2017; Thomke, 2020) and much more work needs to be done, beginning with the role of senior business leaders. To examine their role, we construct and analyze a proprietary dataset of 6,375 experiments on the A/B/n testing platform Optimizely. This unique dataset gives us insights into learning modes, design choices and performance outcomes across industries and firms, which, in turn, leads to the following contributions.

First, we contribute to a burgeoning literature on experimentation in strategy by introducing the study of organizational structure and design choices in experimentation. We find that increasing the hierarchical rank of the most senior person in an organization associates with more learning signals but decreased performance. Furthermore, we find that senior managers' influence may flow through experimental design choices. In particular, seniority in management supports complex tests that associate with a greater chance of significant learning signals; however, such seniority also undermines parallel testing which associates with improved performance. Overall, contrary to the views of practitioners who remain wary of executive influence in testing, senior managers are neither an unambiguous boon or curse to experimentation.

Second, our findings contribute more generally to understanding the limitations that organizational design may place on data-driven decision-making in the digital age (Brynjolfsson and McElheran, 2016; Puranam *et al.*, 2014). Despite sustained scholarly interest in understanding the influence of hierarchy on firm learning and performance, the literature lacks consensus on hierarchy's influence (Damanpour and Aravind, 2012). Moreover, empirical findings are often sensitive to the conceptualization of hierarchy and its context (Bunderson *et al.*, 2016). Indeed, the concept of the "Mendelian" executive, enabled by inexpensive experimentation, suggests that senior managers are well-positioned to lead their firms into a new paradigm for organizational search, overcoming the traditional myopia of hierarchy in the search for new strategy (Levinthal, 2017).

To test this argument, we conceptualize hierarchy in terms of its steepness (Anderson and Brown, 2010), induced by increasing management seniority on an experimentation organization. In particular, we find that hierarchical steepness associates with positive detections in experimentation, albeit at the potential cost of lower performance. This supports views of hierarchy's conservative influence on organizational search, where hierarchies avoid unprofitable projects while simultaneously failing to greenlight potentially profitable ones (Csaszar, 2012; Knudsen and Levinthal, 2007).

Third, we introduce the concept of two distinct learning modes in experimentation—a discovery-driven vs. an optimization approach—and discuss their respective trade-offs in strategic decision-making. Our results demonstrate that a discovery-driven approach, embodied by many simultaneous changes within an experiment, may help generate significant signals which give confidence to organizations about the direction of new strategic path. Nonetheless, this discovery-driven approach is at odds with cause-and-effect learning, which favors small treatments to be allocated across multiple variants. Although this parallel testing strategy may introduce more errors, it can help screen high-performing ideas more effectively (Azevedo *et al.*, 2019). Thus, organizations may choose to toggle between learning modes depending on their experimental objectives—whether it is to validate the choice of a new path via a discovery-driven approach, or to maximize performance along an already chosen path via optimization.

We conclude by noting limitations to the present study and opportunities for future work. First, our study is conducted using a sample of A/B/n tests in the web domain. Despite the strength of our sample in representing A/B/n testing practice across contexts, our findings do not capture experimentation dynamics in non-web settings. Future work could examine the degree to which our findings generalize to other settings, such as offline, physical business experiments (e.g., the testing of non-digital business models), where underlying conditions differ. Second, while our data on the internal hierarchy of testing accounts lends itself well to conceptualizations of the steepness of hierarchy, it does not directly capture other potential constructs of interest, such as cross-relationships (Burton and Obel, 2004) or span of control (Rajan and Wulf, 2006). Here, follow-on research could examine the influence of these alternate mechanisms. Finally, our findings on hierarchy's association with discovery at the potential cost of performance raises interesting questions about mechanisms which we were unable to capture in the present study. For instance, are senior managers taking a longer-term strategic view, wishing to learn about elements of a new strategy while accepting short-term losses in performance? Future research that pairs experimentation choices with long-term performance outcomes would help address this question.

References

Aldrich HE. 1979. *Organizations and Environments*. Prentice Hall: Englewood Cliffs, N.J.

Anderson C, Brown C. 2010. The functions and dysfunctions of hierarchy. *Research in Organizational Behavior* **30**(12): 55–89.

Anderson C, Galinsky AD. 2006. Power, optimism, and risk-taking. *European Journal of Social Psychology* **36**: 511–536.

Andries P, Debackere K, Van Looy B. 2013. Simultaneous Experimentation as a Learning Strategy: Business Model Development Under Uncertainty. *Strategic Entrepreneurship Journal* **7**: 288–310.

Argote L, Miron-Spektor E. 2011. Organizational Learning: From Experience to Knowledge. *Organization Science* **22**(5): 1123–1137.

Azevedo EM *et al.* 2019. *A/B Testing with Fat Tails*. Working Paper.

Billinger S, Stieglitz N, Schumacher TR. 2014. Search on Rugged Landscapes: An Experimental Study. *Organization Science* **25**(1): 93–108.

Brynjolfsson E, McElheran K. 2016. The rapid adoption of data-driven decision-making. *American Economic Review* **106**(5): 133–139.

Bunderson JS, Van der Vegt GS, Cantimur Y, Rink F. 2016. Different Views of Hierarchy and Why They Matter: Hierarchy as Inequality or as Cascading Influence. *Academy of Management Journal* **59**(4): 1265–1289.

Burgelman RA. 1994. Fading Memories: A Process Theory of Strategic Business Exit in Dynamic Environments. *Administrative Science Quarterly* **39**(1): 24.

Burton RM, Obel B. 2004. *Strategic Organizational Diagnosis and Design: The Dynamics of Fit.* Springer: Boston, MA.

Camuffo A, Cordova A, Gambardella A, Spina C. 2019. A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial. *Management Science* (in press).

Carter NM, Gartner WB, Reynolds PD. 1996. Exploring start-up event sequences. *Journal of Business Venturing* **11**(3): 151–166.

Contigiani A, Levinthal DA. 2019. Situating the Construct of Lean Startup: Adjacent "Conversations" and Possible Future Directions. *Industrial and Corporate Change*.

Csaszar FA. 2012. Organizational structure as a determinant of performance: Evidence from mutual funds. *Strategic Management Journal* **33**: 611–632.

Czitrom V. 1999. Teacher's Corner One-Factor-at-a-Time Versus Designed Experiments **53**(2): 126–131.

Dalton, D.R., Todor, W.D., Spendolin, M.J., Fielding, G.J. and Potor LW. 1980. Organization structure and performance. *A critical review. Academy of Management Review* **13**(4): 49–64.

Damanpour F, Aravind D. 2012. *Organizational Structure and Innovation Revisited: From organic to ambidextrous structure. Handbook of Organizational Creativity*, Mumford M (ed). Elsevier Inc.: Oxford, UK.

Dmitriev P, Gupta S, Kim DW, Vaz G. 2017. A Dirty Dozen: Twelve common metric interpretation pitfalls in online controlled experiments. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **Part F1296**: 1427–1436.

Eggers JP, Kaplan S. 2009. Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change. *Organization Science* **20**(2): 461–477.

Fast NJ, Gruenfeld DH, Sivanathan N, Galinsky AD. 2009. Illusory control: A generative force behind power's far-reaching effects. *Psychological Science* **20**: 502–508.

Gans JS, Stern S, Wu J. 2019. Foundations of Entrepreneurial Strategy. *Strategic Management Journal* **40**(5): 736–756.

Garcia-Macia D, Hsieh C-T, Klenow PJ. 2019. How Destructive Is Innovation? *Econometrica* **87**(5): 1507–1541.

Gavetti G, Greve HR, Levinthal DA, Ocasio W. 2012. The Behavioral Theory of the Firm: Assessments and Prospects. *The Academy of Management Annals* **6**: 1–40.

Gavetti G, Levinthal D. 2000. Looking Forward and Looking Backward: Cognitive and Experiential Search. *Administrative Science Quarterly* **45**(1): 113–137.

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* **38**(2): 193–205.

Greene W. 2004. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal* **7**(1): 98–119.

Gruber M, MacMillan IC, Thompson JD. 2008. Look before you leap: Market opportunity identification in emerging technology firms. *Management Science* **54**(9): 1652–1665.

Hage J. 1965. An axiomatic theory of organizations. *Administrative Science Quarterly* **10**: 289–320.

Hall RH, Tolbert PS. 2005. *Organizations: Structures, processes, and outcomes*, 8th ed. Prentice Hall: Upper Saddle River, NJ.

Harvey N, Fischer I. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes* **70**(2): 117–133.

Heckman JJ, Lochner L, Taber C. 1999. Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy. *Fiscal Studies* **20**(1): 25–40.

Hollander S. 1965. *The Sources of Increased Efficiency*.

Kaplan S. 2008. Cognition, capabilities, and incentives: Assessing firm response to the fiber-optic revolution. *Academy of Management Journal* **51**(4): 672–695.

Keum DD, See E. 2017. The Influence of Hierarchy on Idea Generation and Selection in the Innovation Process. *Organization Science* **28**(4): 653–669.

Knight K. 1963. *A Study of Technological Innovation: The Evolution of Digital Computers*. Carnegie Institute of Technology.

Knudsen T, Levinthal DA. 2007. Two Faces of Search: Alternative Generation and Alternative Evaluation. *Organization Science* **18**(1): 39–54.

Kohavi R, Henne RM, Sommerfield D. 2007. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. In *Proceedings of The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007: 959.

Kohavi R, Thomke S. 2017. The Surprising Power of Online Experiments: Getting the Most Out of A/B and Other Controlled Tests. *Harvard Business Review* **95**(5): 74–82.

Koning R, Hasan S, Chatterji A. 2019. *Experimentation and Startup Performance: Evidence from A/B Testing*. HBS Working Paper 20-018.

Lee F, Edmondson AC, Thomke S, Worline M. 2004. The mixed effects of inconsistency on experimentation in organizations. *Organization Science* **15**(3).

Lee S. 2020. *The Myth of the Flat Start-up : Reconsidering the Organizational Structure of Start-ups*. Working Paper.

Lee S, Meyer-Doyle P. 2017. How performance incentives shape individual exploration and exploitation: Evidence from microdata. *Organization Science* **28**(1): 19–38.

Levinthal DA. 1997. Adaptation on Rugged Landscapes. *Management Science* **43**(7): 934–950.

Levinthal DA. 2017. Mendel in the C-Suite: Design and the Evolution of Strategies. *Strategy Science* **2**(4): 282–287.

Levy O. 2005. The influence of top management team attention patterns on global strategic posture of firms. *Journal of Organizational Behavior* **26**(7): 797–819.

Loch CH, Terwiesch C, Thomke S. 2001. Parallel and Sequential Testing of Design Alternatives. *Management Science* **47**(5): 663–678.

March JG. 1991. Exploration and Exploitation in Organizational Learning. *Organization Science* **2**(1): 71–87.

McDonald RM, Eisenhardt KM. 2019. Parallel Play: Startups, Nascent Markets, and Effective Business-model Design. *Administrative Science Quarterly* : 1–41.

Millard A. 1990. *Edison and the Business of Innovation*. John Hopkins University Press: Baltimore, MD.

Mintzberg H. 1978. Patterns in Strategy Formation. *Management Science* **24**(9): 934–948.

Montgomery DC. 2013. *Design and Analysis of Experiments*, 8th ed. Wiley.

Murray F, Tripsas M. 2004. The exploratory processes of entrepreneurial rms: The role of purposeful experimentation. *Advances in Strategic Management* **21**: 45–76.

Ocasio W. 1997. Towards an Attention-Based View of the Firm. *Strategic Management Journal* **18**(S1): 187–206.

Ott TE, Eisenhardt KM, Bingham CB. 2017. Strategy Formation in Entrepreneurial Settings: Past Insights and Future Directions. *Strategic Entrepreneurship Journal* **11**(3): 306–325.

Pekelis L, Walsh D, Johari R. 2015. *The New Stats Engine*. Available at: http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf.

Pich MT, Loch CH, Meyer A De. 2000. On Uncertainty , Ambiguity , and Complexity in Project Management : 7–12.

Puranam P, Alexy O, Reitzig M. 2014. What's "New" About New Forms of Organizing? *Academy of Management Review* **39**(2): 162–180.

Puranam P, Stieglitz N, Osman M, Pillutla MM. 2015. Modelling Bounded Rationality in Organizations: Progress and Prospects. *Academy of Management Annals*. Taylor & Francis **9**(1): 337–392.

Rajan RG, Wulf J. 2006. The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies. *The Review of Economics and Statistics* **88**(4): 759–773.

Reitzig M, Maciejovsky B. 2013. Corporate Hierarchy and Vertical Information Flow Inside the Firm - A Behavioral View. *Strategic Management Journal* **36**: 1979–1999.

Ries E. 2011. The Lean Startup. Crown Business.

Rivkin JW. 2000. Imitation of complex strategies. *Management Science* **46**(6): 824–844.

Rosenbaum PR. 2017. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press: Cambridge, MA.

Scott WR. 1998. *Organizations: Rational, natural and open systems*, 4th ed. Prentice Hall: Upper Saddle River, NJ.

See KE, Morrison EW, Rothman NB, Soll JB. 2011. The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes* **116**(2): 272–285.

Simon HA. 1962. The Architecture of Complexity. In *American Philosophical Society*, 106:

467–482.

Sommer SC, Loch CH. 2004. Selectionism and Learning in Projects with Complexity and Unforeseeable Uncertainty. *Management Science* **50**(10): 1334–1347.

Tannenbaum AS, Kavcic B, Rosner M, Vianello M, Wieser G. 1974. *Hierarchy in organizations*. Jossey-Bass: San Francisco.

Thomke S. 2003. *Experimentation Matters*, 1st ed. Harvard Business School Publishing Corporation: Boston, MA.

Thomke S. 2020. *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Review Press: Boston, MA.

Thomke S, Bell DE. 2001. Sequential Testing in Product Development. *Management Science* **47**(2): 308–323.

Thomke S, Beyersdorfer D. 2018. Booking.com. In *HBS Case No. 9-610-080.*

Thomke S, von Hippel E, Franke R. 1998. Modes of experimentation: an innovation process—and competitive—variable. *Research Policy* **27**(3): 315–332.

Thomke S, Kuemmerle W. 2002. Asset accumulation, interdependence and technological change: Evidence from pharmaceutical drug discovery. *Strategic Management Journal* **23**(7): 619–635.

Thomke SH. 1998. Managing Design in the Experimentation of New Products. *Management Science* **44**(6): 743–762.

Table 1: Descriptive Statistics and Pairwise Correlations (*n* = 6,375).

Variable	Mean	St. Dev.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Positive Statsig	0.107	0.310	0	1	1										
Max Lift	0.103	0.673	-0.915	12.905	0.202	1									
Max Seniority	4.538	1.155	1	6	0.039	-0.026	1								
Variant Count	2.430	0.835	2	8	0.034	0.075	-0.064	1							
Max Variant Complexity	1.355	0.589	1	4	0.027	0.016	0.029	0.02	1						
Mean Variant Complexity	1.319	0.550	1.000	4.000	0.026	0.007	0.033	-0.076	0.967	1					
Duration	4.044	3.845	1	31	0.062	0.016	0.03	0.008	0.019	0.022	1				
Traffic	35.373	383.488	1.001	24,054.730	0.034	0.037	0.016	0.01	0.001	0.003	0.03	1			
Firm Age	33.240	37.569	1	282	0.046	0.021	0	0.019	-0.002	-0.001	0.022	0.011	1		
Employee Count	18,341.170	59,144.170	5	377,757	0.011	0.022	0.095	-0.047	-0.014	-0.014	0.055	0.047	0.349	1	
Technological Integrations	21.934	4.296	0	27	0.02	0.006	0.071	-0.035	-0.048	-0.049	0.01	0.017	-0.027	-0.061	1

Table 2: Hierarchy on Performance. Ordinary least squares (OLS) estimation of cross-sectional data at the experiment level. Robust standard errors clustered at the team level are shown in parentheses and *p*-values are shown in brackets. **p* < 0.10, ***p* < 0.05, ****p* < 0.01.

	ln(Max Lift + 1) (2-1)	Positive Statsig (2-2)
Max Seniority	-0.009^{**}	0.010^{**}
	(0.004)	(0.005)
	[0.016]	[0.047]
Duration	0.002	0.005^{***}
	(0.001)	(0.001)
	[0.165]	[0.0002]
Traffic	0.00000	0.00003^{*}
	(0.00000)	(0.00002)
	[0.364]	[0.076]
Firm Age	0.0002^{*}	0.0005^{**}
	(0.0001)	(0.0002)
	[0.068]	[0.019]
Employee Count	0.00000	-0.00000
	(0.00000)	(0.00000)
	[0.236]	[0.506]
Technological Integrations	0.0004	0.001
	(0.001)	(0.001)
	[0.634]	[0.269]
Industry Fixed Effects	Yes	Yes
Week Fixed Effects	Yes	Yes
R^2	0.0113	0.017
Observations	6,375	6,375

Table 3: Hierarchy on Experiment Design. Ordinary least squares (OLS) estimation of cross-sectional data at the experiment level. Robust standard errors clustered at the team level are shown in parentheses and p-values are shown in brackets. *p < 0.10, **p < 0.05, ***p < 0.01.

	Max Variant Complexity	Variant Count	Mean Variant Complexity
	(3-1)	(3-2)	(3-3)
Max Seniority	0.017^{**}	-0.037^{**}	0.018^{**}
	(0.008)	(0.018)	(0.007)
	[0.036]	[0.038]	[0.018]
Duration	0.003	0.005	0.003
	(0.002)	(0.003)	(0.002)
	[0.126]	[0.145]	[0.105]
Traffic	0.00000	0.00004^{**}	0.00000
	(0.00001)	(0.00002)	(0.00001)
	[0.925]	[0.024]	[0.771]
Firm Age	0.0002	0.001	0.0002
	(0.0003)	(0.001)	(0.0003)
	[0.487]	[0.222]	[0.489]
Employee Count	−0.00000	-0.00000^{*}	-0.00000^{*}
	(0.00000)	(0.00000)	(0.00000)
	[0.106]	[0.076]	[0.067]
Technological Integrations	-0.008^{***}	−0.005	-0.007^{**}
	(0.003)	(0.005)	(0.003)
	[0.009]	[0.245]	[0.012]
Industry Fixed Effects	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes
R^2	0.0149	0.0332	0.0157
Observations	6,375	6,375	6,375

Table 4: Experiment Design on Performance. Ordinary least squares (OLS) estimation of cross-sectional data at the experiment level. Robust standard errors clustered at the team level are shown in parentheses and p-values are shown in brackets. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

	ln(Max Lift + 1)	ln(Max Lift + 1)	Positive Statsig	Positive Statsig	Positive Statsig
	(4-1)	(4-2)	(4-3)	(4-4)	(4-5)
Variant Count	0.038^{***} (0.007) [0.00000]		0.009 (0.007) [0.198]		
Max Variant Complexity		0.014^{*} (0.008) [0.079]		0.019^{**} (0.009) [0.041]	
Mean Variant Complexity					0.021^{**} (0.010) [0.032]
Duration	0.001 (0.001) [0.409]	0.001 (0.001) [0.358]	0.008^{***} (0.002) [0.00001]	0.008^{***} (0.002) [0.00002]	0.008^{***} (0.002) [0.00002]
Traffic	0.00001 (0.00001) [0.148]	0.00001 (0.00001) [0.122]	0.00003^{**} (0.00001) [0.024]	0.00003^{**} (0.00001) [0.023]	0.00003^{**} (0.00001) [0.023]
Team Fixed Effects	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes
R^2	0.264	0.258	0.253	0.254	0.254
Observations	6,375	6,375	6,375	6,375	6,375

Appendix

Table A1: Organizational Level Associations with Learning, Performance, and Experiment Design Choices. Ordinary least squares (OLS) estimation of cross-sectional data at the team level. Robust standard errors clustered at the team level are shown in parentheses and p-values are shown in brackets. *p < 0.10, **p < 0.05, ***p < 0.01.

	Max Lift	Positive Statsig	Max Variant Complexity	Variant Count
	(A1-1)	(A1-2)	(A1-3)	(A1-4)
Max Seniority	-0.046^{**}	0.047^{***}	0.077^{***}	-0.066^{***}
	(0.022)	(0.011)	(0.019)	(0.019)
	[0.039]	[0.00003]	[0.00005]	[0.0005]
Mean Traffic	−0.0001	0.001^{***}	0.001^{***}	0.001^{**}
	(0.0001)	(0.0003)	(0.0004)	(0.0003)
	[0.608]	[0.001]	[0.005]	[0.029]
Mean Duration	0.003	−0.003	-0.019^{***}	−0.001
	(0.008)	(0.004)	(0.006)	(0.005)
	[0.713]	[0.426]	[0.003]	[0.861]
Firm Age	0.001	0.001	−0.0004	0.0001
	(0.001)	(0.0004)	(0.001)	(0.001)
	[0.189]	[0.202]	[0.553]	[0.864]
Employee Count	−0.00000	−0.00000	−0.00000	-0.00000^{**}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
	[0.804]	[0.672]	[0.222]	[0.017]
Technological	0.002	0.003	−0.006	-0.012^{***}
Integrations	(0.004)	(0.003)	(0.005)	(0.005)
	[0.568]	[0.339]	[0.273]	[0.010]
Industry Fixed Effects	Yes	Yes	Yes	Yes
R^2	0.0168	0.0528	0.0444	0.14
Observations	1,101	1,101	1,101	1,101

Table A2: Pre-Experiment Experience. Ordinary least squares (OLS) estimation of cross-sectional data at the experiment level. Robust standard errors clustered at the team level are shown in parentheses and p-values are shown in brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	ln(Max Lift + 1)	Positive Statsig
	(A2-1)	(A2-2)
Max Seniority	-0.009^{**}	0.010^{**}
	(0.004)	(0.005)
	[0.018]	[0.036]
Experimental Experience	-0.00000	-0.00001
	(0.00001)	(0.00001)
	[0.889]	[0.479]
Duration	0.002	0.005^{***}
	(0.001)	(0.001)
	[0.177]	[0.0002]
Traffic	0.00000	0.00003^{*}
	(0.00000)	(0.00002)
	[0.336]	[0.079]
Firm Age	0.0002^{**}	0.0005^{**}
	(0.0001)	(0.0002)
	[0.042]	[0.016]
Employee Count	0.00000	-0.00000^{*}
	(0.00000)	(0.00000)
	[0.185]	[0.072]
Technological Integrations	0.0004	0.001
	(0.001)	(0.001)
	[0.659]	[0.267]
Industry Fixed Effects	Yes	Yes
Week Fixed Effects	Yes	Yes
R^2	0.0112	0.0175
Observations	6,375	6,375

Table A3: Diminishing Returns in Experimentation. Ordinary least squares (OLS) estimation of cross-sectional data at the experiment level. Robust standard errors clustered at the team level are shown in parentheses and p-values are shown in brackets. *p < 0.10, **p < 0.05, ***p < 0.01.

	ln(Max Lift + 1)	Positive Statsig
	(A3-1)	(A3-2)
Max Seniority	−0.007**	0.010**
	(0.004)	(0.005)
	[0.046]	[0.043]
Number Prior Experiments	−0.001	0.0001
	(0.0005)	(0.001)
	[0.120]	[0.890]
Experimental Experience	−0.00000	−0.00001
	(0.00001)	(0.00001)
	[0.989]	[0.483]
Duration	0.001	0.005***
	(0.001)	(0.001)
	[0.291]	[0.0003]
Traffic	0.00000	0.00003*
	(0.00000)	(0.00002)
	[0.226]	[0.077]
Firm Age	0.0002**	0.0005**
	(0.0001)	(0.0002)
	[0.039]	[0.016]
Employee Count	0.00000	−0.00000*
	(0.00000)	(0.00000)
	[0.184]	[0.072]
Technological Integrations	0.001	0.001
	(0.001)	(0.001)
	[0.569]	[0.265]
Industry Fixed Effects	Yes	Yes
Week Fixed Effects	Yes	Yes
R^2	0.0117	0.0175
Observations	6,375	6,375