Experimental Approaches to Strategy and Innovation

Experimental Approaches to Strategy and Innovation

A dissertation presented

by

Sourobh Ghosh

to

The Technology and Operations Management Unit

in partial fulfillment of the requirements

for the degree of

Doctor of Business Administration

in the subject of

Technology and Operations Management

Harvard University

Cambridge, Massachusetts

April 2021

Dissertation Advisor: Andy Wu, Stefan Thomke Sourobh Ghosh

Experimental Approaches to Strategy and Innovation

Abstract

Firms continue to adopt business experimentation to shape product innovation and help form business strategy. In this dissertation, I explore boundary conditions to business experimentation for strategy and innovation practice in three closely related studies. In the first study, I examine how varying access to inexpensive testing influences the way managers search for new opportunities with experiments. Using data from a leading A/B testing platform, I find that greater access to inexpensive testing may inadvertently limit a firm's ability to identify high-performing opportunities. In the second study, I explore how senior managers associate with learning and performance outcomes in experimentation. I find that senior managers' involvement associates with more significant learning signals but smaller performance improvements. In the third study, I analyze how iterative, Agile project management methods that are widely used to manage business experimentation affect new product innovation. In a software development field experiment and follow-on laboratory study, I find that iterative management frameworks cause managers to prioritize short-term value over novelty in products. Overall, this dissertation contributes to a burgeoning literature on experimentation in strategy, innovation, and entrepreneurship. By articulating pertinent boundary conditions to managing experimentation, this work seeks to help managers effectively leverage experimentation to realize its potential benefits for firm innovation and performance.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation exists because of the tireless support of wonderful mentors, friends, and colleagues. I often think of the role they played in helping make this dissertation a reality, leaving me with an overwhelming sense of gratitude. It is with pleasure that I recognize their support here.

I have been truly fortunate to have worked with an exceptional dissertation committee: Andy Wu, Jan Rivkin, Rory McDonald, and Stefan Thomke. Students past and present know what a blessing it is to be advised by any one of these four extraordinary individuals. To have combined their individual talents into a single thesis committee has been a privilege that I have deeply enjoyed these past few years. Andy Wu uniquely challenged me to realize my full potential as a doctoral student. Andy's dedication to developing students and helping them hone their professional craft is unparalleled, and I am better for it. I have often heard that meeting one's intellectual heroes is an intimidating experience; I am happy to report that in my experience with Jan Rivkin, this could not be further from the truth. Jan's intellect and empathy give him an uncanny ability to get to the crux of any issue in no time, which sharpened my thinking. Rory McDonald was my go-to reference on any literature, helping me translate my informal ideas into the language of theoretical constructs. In addition, I deeply valued the wisdom that Rory brought to our wide-ranging conversations on academic life. Finally, I thank Stefan Thomke for being the first person at HBS to take an interest in me—back while I was still an overzealous engineering undergraduate student. Stefan has been the throughline for my entire experience at Harvard; his

attention and humor have made this a better experience for me.

I must recognize the support of a robust academic community at HBS across levels. The TOM and Strategy units provided an excellent environment for me to develop and grow as a scholar. I am forever grateful to my doctoral student colleagues, past and present, for being a constant source of inspiration and support. A few who warrant special mention include: Ryan Allen, Maya Balakrishnan, Hayley Blunden, Yo-Jud Cheng, MoonSoo Choi, John Cromwell, Tommy Pan Fang, Chris Fulton, Cheng Gao, Paul Green, Grace Gu, Adi Gupta, Raha Imanirad, Olivia Jung, Yosub Jung, Do Yoon Kim, Ohchan Kwon, Michael Anne Kyle, Frank Nagle, Ashley Palmarozzo, James Sappenfield, Peter Scoblic, Lumumba Seegars, Lauren Taylor, Mike Teodorescu, Ehsan Valavi, Kala Viswanathan, and Daniel Yue. In addition, I thank the members of the Doctoral Programs Office for their help—no request is too inane or cumbersome for this resourceful group of people. Finally, I have been exceedingly fortunate that many HBS faculty outside my committee took a special interest in my work. Faculty who inspired me to consider new perspectives and dig deeper in my work include: Ethan Bernstein, John Beshears, Iav Bojinov, Ryan Buell, Chiara Farronato, Francesca Gino, Shane Greenstein, Rem Koning, Alan MacCormack, Lakshmi Ramarajan, Mike Tushman, Eric Van den Steen, Dennis Yao, and Feng Zhu.

The research presented in this dissertation would not have been possible without the support of several intellectually curious, diligent, and engaged research partners. A field experiment to evaluate Agile project management's impact on innovation would never have occurred without the support of Google. I owe special thanks to my friend, Max Bileschi, in addition to Mark Bouchard, Todd Burner, and Leah Worbs-Lunde, for making our vision of a field experiment within a software development hackathon a reality. This dissertation also presents unique, proprietary data on digital experimentation, which is made possible by the support of Optimizely. Identifying, cleaning, and preparing this data for rigorous academic study presented unique challenges, all of which were overcome with the continued support of Hazjier Pourkhalkhali and Charles Pensig. Research and administrative staff

across HBS also played significant roles in advancing my research; a few to note include: Alain Bonacossa, Andrew Falzone, Andrew Marder, Kathryn O'Brien, Karin Parodi, Sam Snyder, and Ista Zahn. In addition, I have had the pleasure of working with four remarkable and talented research assistants: Emily Chen, Jaeho Kim, Kiyeon Lee, and Jia Yi Lim.

Two gentlemen in particular deserve special recognition for introducing me to the stimulating world of academic research: Kemper Lewis and Warren Seering. They were among the first to recognize my potential as a researcher while selflessly supporting my transition from engineering design research to management research. Their thoughtful approach to advising should be a model for all.

I gratefully acknowledge the HBS Doctoral Fellowship and the National Science Foundation Graduate Research Fellowship for financially supporting my research and education at HBS. I also appreciate additional financial support for my research from the HBS Division of Research and Faculty Development and the Kauffman Foundation.

Finally, this work would not exist without the unconditional support of my friends and family. There is an old adage in graduate school: a good dissertation is a "done" dissertation. I would posit the following corollary: a dissertation is "done" at the pleasure of the author's closest friends and family. Friends past and present—Adam, Alison, Anders, Christine, Dan, Kate, Katie, Lee, and Nick—provided the cheer necessary to see through the program. Of course, the driving force behind my education—and any of my other endeavors—is my family. My father, Amit, instilled in me a sense of striving for the highest standards. My mother, Ranu, provided constant, reassuring emotional support for the most challenging days of my education. My brother, Sid, has always set a wonderful example for me, proudly serving as my first defacto mentor going back to grade school days. Finally, I owe special thanks to my partner, Raashmi, whose support, patience, and wit have made the last several years so much more fun. I dedicate this work to my family—any success that I enjoy is made sweeter knowing that I get to share it with you. Thank you.

Chapter 1

Introduction

This dissertation considers how managers use business experimentation to shape product innovation and help form business strategy. In recent years, there has been an explosion in interest in business experimentation as a method to manage entrepreneurship and innovation (Bennett and Chatterji 2019, Gans et al. 2019, Contigiani and Levinthal 2019, Felin et al. 2019, Bocken and Snihur 2020, Leatherbee and Katila 2020). For instance, frameworks such as the Lean Startup method have been widely adopted by entrepreneurs, who have used its principles to guide many new venture launches and product innovations (Blank 2003, Ries 2011). In addition to entrepreneurial firms, established corporations, including those without a legacy in industrial R&D or experimentation practice, have adopted experimentation as a methodology to manage their in-house innovation efforts with varying levels of success (Thomke 2020, Gupta et al. 2019, Kohavi et al. 2020). In response to this growing movement in practice, there has been increased focus on experimentation as a topic of academic inquiry within the fields of strategy and entrepreneurship (Camuffo et al. 2020a, Koning et al. 2019, Pillai et al. 2020, Agrawal et al. 2021).

Even within the boundaries of the management sciences, the current interest in experimentation in strategy and entrepreneurship has intellectual antecedents across several adjacent fields, such as innovation management (Thomke 2003), operations management

(Bohn and Lapré 2011), and entrepreneurial finance (Kerr et al. 2014), to name a few examples.[1] While various subfields of management will conceptualize experimentation in slightly different ways, they agree upon the fundamentals. In particular, a business experiment allows decision-makers to reduce the uncertainty of pursuing a new business idea.[2]

Various disciplines also agree on the benefits of business experimentation, which are numerous and multi-faceted. First and foremost, each business experiment is an opportunity for the firm to learn (Thomke 2003). Pursuing new business ideas is inherently risky and uncertain. Learning through experimentation is especially helpful when the firm cannot deduce the value of a new business idea from theory or prior experience alone (Rosenberg 1994, Pillai et al. 2020). In addition, a process of experimentation can help managers discover new and unanticipated business ideas (Eisenhardt and Bingham 2017). By prioritizing learning through experimentation, firms enjoy tangible performance benefits. Firms which adopt experimentation perform better than their peers (Koning et al. 2019) as they are more likely to learn about and avoid false positive ideas that may look promising but are ultimately unprofitable to pursue (Camuffo et al. 2020a). In addition, frequent experimentation increases managers' chance of identifying and validating outlier ideas which provide superlative performance returns for the firm (Azevedo et al. 2019).

1.1 Experimental Approaches to Strategy and Innovation: A Research Agenda

Despite a well-established cross-disciplinary understanding of experimentation's benefits, many firms have only recently begun to adopt experimentation in strategy and entrepreneurship practice (Koning et al. 2019, Thomke 2020). Why might help explain this trend, and what implications does this have for scholarship in strategic management? While business experimentation has always been a desirable input to firm innovation processes, it has histori-

[1] For a more extensive review, see Contigiani and Levinthal (2019).

[2] I follow Azevedo et al. (2019) and refer to experimental treatments as "ideas," which may encompass new products, services, algorithm improvements, and other initiatives intended to improve business performance.

cally been cost prohibitive for organizations to conduct. Advances in information technology, among other factors, have made the cost to organizations to design and implement experiments cheaper than at any other point in business history (Thomke 2003). As a result, business experimentation can now be used to test and inform decisions of firm-wide consequence. In particular, firms can now evaluate the performance of interdependent choices with experiments, where interdependence is the quality that makes a decision strategic (Eisenhardt and Bingham 2017, Van den Steen 2017, Leiblein et al. 2018). We will explore examples of experimentation's impact on strategic, interdependent choices—such as the decision to enter a new market—in Chapter 2. In short, the use of inexpensive experimentation changes the way practitioners think about and form their strategy.

I argue that the use of inexpensive experimentation to test strategic decisions alters strategy and entrepreneurship practice in at least three fundamental ways—through *inexpensive learning from customers*, *partial commitment*, and *rapid iteration*. Together, these three characteristics define what I call "experimental approaches to strategy and innovation." Below, I detail each of the three characteristics and their potential implications for strategic management practice.

The first meaningful difference for strategy is from *inexpensive learning from customers*. Learning is a key function in the formulation of strategy as it allows an organization's members to reinforce actions that improve the firm's performance (Gavetti et al. 2012, Puranam et al. 2015). Traditionally, learning has been a noisy process in strategy formulation (Cyert and March 1963). While a firm's customers help determine the performance of a strategy (Christensen and Bower 1996, Felin et al. 2019), firms are typically constrained in their ability to try new strategies with customers without committing significant resources. As a result, managers often use imperfect proxies to determine the performance of a strategy, such as simulations or mental representations (Csaszar 2018). With inexpensive learning from experimentation, firms can cheaply and efficiently solicit performance feedback directly from customers instead of simulating their response to a new strategy.

With direct feedback from customers, experimenting firms have an evidence-based mechanism to "pivot" or change directions in strategy (Ries 2011, McDonald and Eisenhardt 2019), embodying the second characteristic of *partial commitment*. Commitment, or the "tendency toward. . . persistence" in strategy (Ghemawat 1991, p. 13), provides incentives to organizational members for their effort and leads to more efficient coordination of organization action, among other benefits (Van den Steen 2017). In contrast, experimentation challenges conventional wisdom on commitment by compelling decision-makers to remain engaged in one course of action while simultaneously exploring new opportunities (Levinthal 2017), thereby encouraging *partial* commitment (Gans et al. 2019). While some scholars criticize experimentation for potential opportunity costs incurred from partial commitment (Gans et al. 2019, Felin et al. 2019), other scholars have provided evidence for the benefits of a more flexible approach to strategy, especially in volatile environments in which experimentation is commonly used (Van den Steen 2017).[3] Under a regime of experimentation, strategic reorientations or "pivots" are achieved not by a single decision, but by incrementally adding (removing) profitable (unprofitable) elements over time (Kirtley and O'Mahony 2020). The cumulative benefit of these evidence-based incremental changes is often argued to match or exceed those of full-commitment strategic reorientations, which have a high likelihood of failure (Levinthal 2017, Thomke 2020).

Finally, with inexpensive learning and partial commitment, a firm's products, services, and strategic choices remain constantly evolving—inviting the third characteristic of experimental approaches to strategy and innovation, *rapid iteration*. Iterative frameworks such as Agile project management have evolved to help manage the experimental process within organizations (Rigby et al. 2016a). While practitioners often highlight how rapid iteration improves the speed of innovation in volatile, fast-changing business environments

[3] Van den Steen (2017, p. 2623) notes that "in a volatile environment, strategy should be built more around internal factors, such as capabilities and resources, that are under the control of the organization and can thus be kept stable, rather than around products or needs that may change quickly. This is clear in high-tech industries where firms often build their strategy around their capabilities and broad market needs, rather than around specific products or solutions."

(Eisenmann et al. 2013), rapid iteration also provides a mechanism for organizations to revisit and potentially reprioritize their strategic objectives (Rigby et al. 2016a, Relihan 2018). In effect, this enables strategic reorientations necessary for the firm to survive and thrive. We will explore some of these arguments in further detail in Chapter 4.

Together, these characteristics help define what makes experimentation in strategy and innovation practice unique and worthy of additional study. Burgeoning literature in this domain has largely focused on the benefits of adopting experimentation (Koning et al. 2019, Camuffo et al. 2020a). Nonetheless, practical guidance for managers who have already adopted experimentation remains limited. This is not due to a lack of scholarly interest; rather, it is often due to a lack of reliable data on how experimentation is managed within firms (Kohavi et al. 2020). Accordingly, a major contribution of this thesis is to introduce novel data sources—observational and experimental—to help unpack how experimentation is managed within firms.

Despite the continued adoption of experimental approaches to strategy and innovation in practice, we lack a scholarly understanding of how managers run experiments within their organizations and what effects this may on a firm's broader innovation and performance objectives. For instance, while experimentation is well-suited to optimize firm performance in the short-run, how does it influence the pursuit of broader strategic objectives such as novelty in innovation? This thesis explores potential boundary conditions to managing business experimentation within organizations at scale—a topic which remains underexplored in existing literature (Thomke 2020). Armed with this understanding, managers may more effectively leverage experimentation to realize its potential benefits for firm innovation and performance.

1.2 Thesis Outline

I now briefly summarize three studies which comprise this dissertation. These studies represent an initial attempt to advance our understanding of experimental approaches to strategy—with chapters 2 and 3 examining managing under conditions of inexpensive learn-

ing from customers, and chapter 4 focusing on the characteristic of rapid iteration. Future work should continue to unpack the effects of other relevant characteristics to experimentation within organizations.

In Chapter 2, I examine the influence of the first characteristic of inexpensive learning from customers. As the cost of experimentation continues to decrease, we might expect that firms would improve in their ability to evaluate new strategic alternatives with cheaper customer feedback—helping them identify paths that yield the highest performance. Using data from a leading web experimentation platform, what I find is the opposite of what many practitioners and scholars might expect—that is, I find that access to cheap testing may inadvertently harm a firm's ability to identify high-performing returns. This occurs because as the cost of testing falls, the cognition necessary to design interdependent changes becomes relatively more expensive to members within the firm. This is pertinent as I also show that interdependent changes in experimentation associate with a greater chance of performance breakthrough and a *decreased* chance of performance failure. My findings help demonstrate that access to inexpensive experimentation cannot replace the benefits of managerial cognition and the "offline" search and evaluation of alternatives in strategy. Nonetheless, my findings also suggest a preliminary mechanism by which organizations may expand their search efforts—limiting access to testing. In doing so, organizations can help managers focus on which alternatives to pursue and which not to pursue (Porter 1996) when experimenting to help form strategy.

Access to inexpensive experimentation also raises a tricky question about organizational structure and management—if more strategic decisions are being made with the help of business experiments, then what is the role of senior managers? In Chapter 3, with Stefan Thomke and Hazjier Pourkhalkhali, I explore how increasing seniority in management associates with learning and performance outcomes in experimentation. Our findings suggest that senior management's association is mixed: While senior managers' involvement associates with more complex experiments that create more significant learning signals, their

involvement also associates with smaller performance improvements. Together, these results contribute to our understanding of how to design data-driven organizations, including the benefits and consequences of involving senior leaders who may support experimentation with additional resources while undermining it with potential decision-making biases. This chapter also maps senior leaders' association on experimentation to specific experiment design choices. This analysis offers insight into how managers can modulate organizational search and performance outcomes through experimental design.

Finally, in Chapter 4, I study how iterative frameworks designed to manage the business experimentation process may influence product innovation. Popular practitioner frameworks like Agile management suggest that organizations can adopt an iterative approach of frequent meetings to prioritize between these goals, a practice I refer to as *iterative coordination*. With Andy Wu and the information technology firm Google, I embed a field experiment within a hackathon software development competition to identify the effect of iterative coordination on product innovation. We find that, although iterative coordination leads firms to develop products that are judged to be more valuable, these products are simultaneously less novel. Furthermore, by tracking minute-by-minute changes in software source code, we find that iteratively coordinating firms favor knowledge integration at the cost of in-depth, specialized knowledge creation by their members. A follow-on laboratory study documents that increasing the frequency and opportunities to reprioritize goals in iterative coordination meetings reinforces value and integration, while reducing novelty and specialization. Overall, this chapter helps illustrate how iterative management frameworks may cause managers to implicitly favor value over innovation's definitional characteristic of novelty—suggesting that managers must be more intentional about introducing novelty into experimental processes. Furthermore, we introduce new methodology for study innovation process—namely through software code version-tracking and the use of hackathons to study new forms of organizing.[4]

[4] For a technical review of new forms of organizing, see Puranam et al. (2014).

Chapter 2

Think Before You Act: The Unintended Consequences of Inexpensive Business Experimentation

Sourobh Ghosh

Abstract. Scholars and practitioners recommend the use of inexpensive business experiments to evaluate new and uncertain strategic alternatives. While current thinking recommends that strategic alternatives be tested as a series of *independent* changes across many experiments, this contradicts scholarly understanding of strategy as the formulation of *interdependent* activities that when combined together drive superior performance. To evaluate the tension between testing interdependent and independent changes in experimentation, I first conduct an exploratory analysis of 31,716 business experiments run on using the web experimentation platform, Optimizely. Contrary to popular wisdom, not only does testing a larger set of interdependent changes in an experiment associate with breakthrough performance, it also associates with reduced performance failure. Despite these benefits, I find that a plurality of experiments feature little to no interdependent change.

To explain why firms vary in whether they test interdependent changes, I develop and test theory for how access to testing resources influences the design of interdependent changes in experimentation. I find that greater access to testing leads

firms to test experiments with fewer interdependent changes. In contrast, when access to testing is limited, I find that firms *increase* interdependent changes per experiment. This suggests a potential solution to alleviate an organization's cognitive limits when experimenting: by restricting access to testing resources, firms can focus on testing interdependent activities that have the potential to deliver outstanding performance. Overall, my findings demonstrate the underappreciated value of interdependent changes and its performance benefits for business experimentation.

2.1 Introduction

Business experimentation continues to transform the way firms form their strategy (Levinthal 2017, Andries et al. 2013, Gans et al. 2019). In particular, the decreasing cost of experimentation makes it easier to evaluate strategic alternatives with uncertain returns (Koning et al. 2019, Kerr et al. 2014). To reduce the uncertainty associated with pursuing a new alternative, current thinking by scholars and practitioners advises decision-makers to break up a new alternative into smaller, *independent* changes to be tested individually (Ries 2011, Thomke 2003). For instance, for a discount fast fashion retailer considering whether to introduce luxury goods, it may begin by testing the independent change of introducing marketing copy emphasizing the high-end features of their current products. By testing one change at a time, firms can improve their ability to learn cause-and-effect (Camuffo et al. 2020a). Furthermore, it is argued that testing changes independently not only avoids potential performance losses from coupled changes, but it may also help identify significant performance improvements (Camuffo et al. 2020b, Kohavi et al. 2020, Thomke 2020).

Nonetheless, the prescription to test changes independently of one another in business experimentation contradicts a core property of strategy: how *interdependent* activities combine and reinforce one another to drive superior performance (Siggelkow 2002, Leiblein et al. 2018, Eisenhardt and Bingham 2017). For instance, consider an online platform specializing in hotel bookings evaluating the strategic alternative to become an all-purpose travel platform. Should they expand their offerings and provide flight booking services as well? To appropriately evaluate this alternative via an experiment, the platform must not only

alter the visual design of its website, but it must develop complementary activities such as back-end algorithms that serve users with accurate information on available flights (Thomke and Beyersdorfer 2018). This example illustrates the well-documented importance of interdependent activities to firm strategy (Van den Steen 2017, Leiblein et al. 2018, Eisenhardt and Bingham 2017). Interdependent activities are theorized to act together to drive superior performance (Gavetti and Levinthal 2000). As existing literature has generally modeled business experimentation as a process of evaluating independent changes (Ewens et al. 2018, Azevedo et al. 2019, Shelef et al. 2020), there has been a lack of scholarly attention as to how interdependent changes are generated and evaluated in business experimentation. This suggests a need to develop new theory to explain variation in the quality of interdependent changes and their performance implications in business experimentation.

In this paper, I develop and test theory for how access to testing resources influences the design of interdependent changes in experimentation. In particular, I argue that additional resources for testing leads firms to reduce interdependent changes in experimentation. As firms receive more testing resources, the action of running an additional experiment becomes relatively cheaper to a firm than the cognition necessary to design and build interdependent changes in that experiment. Consequently, firms reduce their investment in the cognition that is necessary to guide interdependent changes to test in a business experiment. By reducing cognition and the pursuit of interdependent changes in an individual experiment, firms may limit their ability to identify a high-performing alternative.

I test my arguments in two parts: first, with an exploratory analysis of the association of interdependent changes on experimental performance; and second, by examining the effect of testing resources on the design of interdependent changes in experimentation. Using data on 31,716 business experiments run on the web experimentation platform, Optimizely, I examine how testing interdependent versus independent changes associate with experimental performance. I find that greater interdependent change associates with an increased chance of breakthrough performance. Furthermore, contrary to popular wisdom that inter-

dependent change may trigger performance failures (Kohavi et al. 2020), I find that greater interdependent change in experimentation also associates with a *reduced* chance of performance failure. This finding cannot be explained due to the relative uncertainty of change alone, which would increase the chance of both success and failure (Fleming 2001). Despite the performance benefits of interdependent change, I find that a plurality of experiments test relatively independent changes.

Next, to explain why firms test independent as opposed to interdependent changes, I combine Optimizely's data on business experiments with a two-year panel of website traffic data to produce quantitative estimates on the effects of additional testing resources on interdependent change. Consistent with my theorizing, I find that cheaper testing leads firms to run more experiments, but that these experiments feature fewer interdependent changes. Examining cognitive mechanisms, I find that cheap testing not only leads firms to reduce variation prior to testing, but they become less selective about which alternatives they wish to test. On the other hand, estimating the effects of a natural experiment that quasi-exogenously makes testing more expensive, I find that firms *increase* interdependent changes per experiment. This suggests a potential solution to alleviate the issue of an organization's cognitive limits when experimenting (Gavetti and Menon 2016): by restricting resources for testing, firms can focus on testing interdependent changes that may deliver outstanding performance.

My findings contribute to the cross-disciplinary literature on business experimentation in three ways. First, I contribute to a burgeoning literature on experimentation in strategy by introducing a framework for scarce resources for cognition and action in business experimentation. This framework guides my analysis of how access to testing resources can shape how organizations search for opportunities. Second, my findings illustrate the value of interdependent change in experimentation, helping organizations unlock long-tail performance potential in experimentation while mitigating the chance of failure. Finally, my findings illustrate how scarce testing resources can shape interdependent change in exper-

imentation. In particular, additional resources for testing reduce the intensity with which firms evaluate their alternatives offline, which may hamper an organization's ability to test different alternatives (Gavetti and Menon 2016). Nonetheless, my findings suggest a mechanism by which organizations may expand their search efforts—limiting access to testing. In doing so, organizations can help managers focus on which alternatives to pursue and which not to pursue (Porter 1996), helping them appreciate the true opportunity cost of running a business experiment (Gans et al. 2019).

This chapter is organized as follows. In Section 2.2, I develop a theoretical framework for scarce resources for cognition and action in experimentation, guiding the development of hypotheses on the effects of testing resources on interdependent change. In Section 2.3, I discuss the empirical context of A/B testing and its relevant features for the study of experimentation in the formation of strategy. In Section 2.4, I present an exploratory analysis that maps the association between interdependent change and performance in experimentation. In Section 2.5, I present my main analysis which studies the effect of testing resources on interdependent change in experimentation. Finally, I summarize contributions of this study in Section 2.6.

2.2 Theoretical Development

Organizations have a long and storied history of using experiments to generate and evaluate new alternatives. Innovations across contexts, from solutions to scurvy to the bestselling Post-it note, were identified and developed with the help of experiments (Bohn and Lapré 2011, Thomke 2003). Today, with the proliferation of cheap testing technology, business experimentation is now available to test decisions of strategic consequence. For instance, should an organization pursue a new product, business model, or technology? Rather than "bet the company" on a new path with uncertain returns, an organization can devise controlled trials where the performance of a new path is tested against current practice (Levinthal 2017). This approach yields tangible benefits for the quality of strategic decision-making, helping organizations avoid "false positives" where new initiatives that look initially promising are

found to be ultimately detrimental to firm performance (Camuffo et al. 2020a). Furthermore, experimentation can help organizations secure high performing returns over time (Kohavi et al. 2020). In particular, experimentation can help organizations screen ideas from the long tail of outstanding performance opportunities (Azevedo et al. 2019).

Business experiments help form strategy via a two-part process of: 1) cognition to guide which decisions to test, followed by 2) action to test these decisions. Cognition allows firms to take a broad view of available activities and their interdependencies, whereas action allows firms to learn about the performance of a set of chosen activities (Gavetti and Rivkin 2007, Ott et al. 2017). Relative to other approaches for strategy formation, experimentation is unique in that it balances the relative advantages and disadvantages of cognition-driven strategy with action-driven strategy. Action-driven methods of strategy formation, such as bricolage (Baker and Nelson 2005), improvisation (Miner et al. 2001), and trial-and-error learning (Bingham and Davis 2012), prioritize information from the external environment to determine the performance of a strategic alternative. While these approaches may be effective at screening decisions for their performance, the organization is often left with an incomplete understanding of *how* activities fit with each other to contribute to performance (Ott et al. 2017). In contrast, cognition-driven approaches to strategy seek to establish a broad understanding of how a firm's interdependent activities may reinforce each other to drive superior performance (Siggelkow 2002, Ott et al. 2017, Eisenhardt and Bingham 2017). Nonetheless, a purely cognition-driven approach absent some form of interim feedback from the external environment is often doomed to fail. In contrast, an experimental approach to strategy formation offers a middle ground between cognition and action (Levinthal 2017), where the organization can vary strategic alternatives according to cognition and its benefits of an understanding of strategy's interdependent connections with the performance feedback benefits of action (Ott et al. 2017).

Despite the importance of both cognition and action, practitioners have celebrated the use of experiments to become data-driven, privileging action and the testing of *indepen-*

dent rather than *interdependent* changes in experimentation. Skeptical of the use of intuition and judgment to drive decision-making, organizations increasingly use business experiments to evaluate small, independent decisions (Thomke 2020). Here, practitioners regard rapid testing on independent changes as a virtue, helping drive unanticipated gains. For instance, Google famously ran an experiment testing 41 shades of blue to decide which to use in a navigation bar. The subsequent performance improvements from this experiment helped inspire similar experiments at peer firms, such as Microsoft's Bing platform (Kohavi et al. 2020, p. 16). Here, the decision of which color to use on a navigation bar is independent of other contemporaneous business decisions and can be tested rapidly. In these and similar cases, organizations use a Darwinian approach of random variation to make several independent changes (Levinthal 2017). Under this approach of random variation, the organization does not have a guiding theory or framework inspiring action; rather, the organization chooses to test an independent decision such as a color change and then to learn from the test. In other words, rather than using cognition to guide action, organizations often use action in experimentation to guide subsequent cognition. This contradicts the classical notion of a business strategy experiment being a cognition-forward process where combinations of interdependent activities are intentionally chosen and designed, followed by action to test such alternatives (Levinthal 2017, Pillai et al. 2020).

Why might organizations favor testing independent changes instead of interdependent changes in experimentation? Here, it is helpful to examine the resources available for cognition and action in business experimentation. Scholars across domains have suggested that organizational resources may influence the design and implementation of business experiments (Thomke 2003, Pillai et al. 2020). Nonetheless, little is known about how specific resources impact the design of experiments in strategy and innovation. Below, I develop a framework for scarce resources for cognition and action in business experimentation. This framework shall guide the development of hypotheses regarding how the cost of testing influences interdependent change in experimentation.

2.2.1 Cognition and Action Resources in Business Experimentation

Business experiments require a collection of scarce resources from across an organization, including those for the processes of cognition and action. These processes, and the resources that support them, are depicted in in Table 2.1.

In the cognition phase, organizations design and build treatments to test. Decision-makers begin this phase by generating and evaluating potential alternatives to test (Knudsen and Levinthal 2007, Ulrich et al. 2020). Each alternative is composed of a set of individual activities. Because all possible combinations of activities cannot be feasibly tested (Rivkin 2000, Levinthal 2017), the organization must evaluate and select which alternatives to test according to beliefs about which subset of activities will yield the highest performance (Ewens et al. 2018, Camuffo et al. 2020a, Gavetti and Levinthal 2000). These beliefs may be informed by a variety of sources, such as mental models (Gary and Wood 2011), frameworks (Csaszar and Levinthal 2016), analogies (Gavetti et al. 2005), theory (Felin and Zenger 2009), and other forms of entrepreneurial judgment (Klein 2008).[5] The purpose of these cognitive efforts is to take a broad view of a firm's interdependent choices and to find the most promising subset of activities to design and build as a potential strategic alternative. The alternative is then built in preparation for running an experiment.

[5] I consider both static knowledge and information processing to provide scarce cognitive resources for the organization (Walsh 1995, Helfat and Peteraf 2015). In this chapter, I abstract away from this distinction and focus on how the cost of cognition changes relative to action in the formation of strategy.

Table 2.1: **Conceptual Framework for Scarce Experimental Resources by Experimental Phase**

Experimental Phase	**Cognition**	**Action**
Source of information	Mental models, representations, analogies	Customers
Phase Processes	Generation of new alternatives Cognitive evaluation	Selection among alternatives
Phase Outcomes	Interdependent Change	Number of experimental trials
Example resources	Engineering expertise (Kohavi et al. 2020) Managerial attention (Ganz 2020, Ghosh et al. 2020) Individual cognitive ability (Csaszar and Laureiro-Martínez 2018) Culture (Thomke 2020)	Customers (Azevedo et al. 2019, Ries 2011) Testing tools (Koning et al. 2019, Thomke 2020) Data/analytics infrastructure (Gupta et al. 2019)

In the action phase, organizations run the experiment to test with customers. Here, customers provide feedback on the new strategic alternative (Ries 2011, Felin et al. 2019, Thomke 2020). The organization can then learn what effect the new alternative has on performance as it is measured relative to current practice. With this information, the organization then selects whether to commit to the new alternative or stick to the baseline of existing practice (Levinthal 2017).

Ultimately, organizations have limited resources for cognition and action in experimentation. Examples of limited cognitive resources include engineering expertise to build complex treatments (Kohavi et al. 2020), executive attention to motivate broader, exploratory changes (Ganz 2020, Ghosh et al. 2020), and individual cognitive capabilities to foresee the interdependencies relevant to a proposed strategic alternative (Csaszar 2018). Other, more subtle cognitive resources include an organization's experimentation culture—which may help an organization unlock latent cognitive resources for experimentation (Lee et al. 2004, Thomke 2020, Meyer et al. 2019).[6] Together, these resources may influence the quality of interdependent change tested in a business experiment. For instance, for an e-commerce retailer to test the activities of whether a new layout for a product landing page and a personalized experience would improve performance, an organization may choose to implement user experience changes and a new personalized recommendation engine that recommends products based on one's search history. To implement this treatment, the organization requires scarce engineering resources to implement the algorithms necessary for the recommendation engine (Kohavi et al. 2020). Accordingly, an observable outcome of the cognition phase is the scope of interdependent change designed in an experiment.

While resources for experimental cognition remain constrained, organizations may benefit from more plentiful resources for experimental action. Scholarly observations about the decreasing cost of experimentation are often premised on the assumption of the decreasing

[6] Culture is an important determinant for firm experimentation (Thomke 2020), where employees must feel empowered to suggest changes to test (Lee et al. 2004).

cost of acquiring feedback information to help an organization learn (Thomke 2003).[7] A variety of resources are required to help an organization collect and process feedback, such as testing tools (Koning et al. 2019, Thomke 2020) and data analytics infrastructure (Gupta et al. 2019). Ultimately, the source of feedback information is provided by customers (Felin et al. 2019, Ries 2011). Traditionally, access to customer feedback has been a key challenge for firms, limiting their ability to learn from the business environment (Clough et al. 2019, Ries 2011). Nonetheless, digitization, among other factors, has enabled firms to reach customers who can provide feedback on a new strategic alternative more effectively (Brynjolfsson and McElheran 2016). Together, the decreasing cost of testing technology and increased access to customers reduces the cost to firms to acquire feedback from customers via additional experiments. Here, an observable outcome of the action phase is the number of experimental trials that can be run with customers.

As an organization benefits from cheaper action resources, the cost of cognition relative to action increases, *ceteris paribus*. Given the role of cognition in helping firms generate and evaluate interdependent activities (Helfat and Peteraf 2015, Baron 2004, Gavetti and Menon 2016), this has important implications for the design of interdependent decisions to test in experimentation. Below, I hypothesize on the effect that action resources have in shaping interdependent changes in experimentation.

2.2.2 Action Resources and Interdependent Change in Experimentation

As organizations receive additional action resources, they are more likely to run additional experimental trials. Additional action resources such as access to customers reduces the cost of experimental trials (Azevedo et al. 2019). Evidence across domains suggests that

[7] For instance, Thomke (2003) documents how decreasing computational costs made experimentation accessible in various industrial R&D settings, such as in pharmaceuticals and the automotive industry. With cheap simulation technologies that would simulate the feedback provided by costly physical prototypes, these firms could experiment with new designs cheaply and efficiently. Note that in my framework for business experimentation, feedback information must be furnished by customers. As a result, I classify simulation as experimental cognition rather than action (see Table 2.1). This classification is consistent with perspectives of simulation as a mode of cognition, which is a simplified representation of the external environment (Gavetti and Levinthal 2000, Ott et al. 2017). For instance, in crash simulation testing, finite element analysis is governed by partial differential equations which are simplified representations of real-world physical phenomena.

with cheaper experimental trials, organizations will elect to run more experiments (Bohn and Lapré 2011). For instance, in entrepreneurial finance, cheaper trials enable investors to spread the risk of initial investments (Ewens et al. 2018), while in product development cheap trials are leveraged to eliminate errors in new products (Thomke 1998). In strategy, as the cost of experimental trials decreases, decision-makers prefer to run additional trials to help reduce causal ambiguity, or the uncertainty surrounding cause-and-effect relationships in strategy (Mosakowski 1997). With each additional trial, decision-makers can reduce causal ambiguity by varying decisions one-at-a-time rather than in broader, interdependent combinations of choices where underlying causal relationships may be poorly understood (Camuffo et al. 2020a).

While helping mitigate causal ambiguity, additional resources for experimental action simultaneously exacerbates the problem of scarce cognitive resources, which may limit interdependent changes in experimentation. Despite the influence of additional action resources in driving additional experimental trials, cognitive resources remain fixed in the short-run (Helfat and Peteraf 2015).[8] As a result, fewer cognitive resources can be devoted per experimental trial. With fewer cognitive resources to devote per experimental trial, firms limit their ability to design and build interdependent sets of activities that differ from the existing practice (Gavetti and Menon 2016). This occurs since an organization has fewer resources to build and evaluate new and unique alternatives prior to running a test (Csaszar and Laureiro-Martínez 2018). Thus, while action resources increase, a desire to mitigate causal ambiguity and limited resources for cognition limit interdependent changes per experiment.

On the other hand, a decrease in action resources may stimulate greater interdependent change per experimental trial. With reduced action resources, firms cannot run as many experimental trials, thereby reducing the data they are able to collect from customer

[8] Helfat and Peteraf (2015) note that ways to extend individual managerial cognitive capabilities remains an ongoing area of research. Suggested interventions include strategy coursework to increase the breadth of representations that individual managers use to design and evaluate strategic alternatives (Csaszar and Laureiro-Martínez 2018, pg. 527). Nonetheless, these and related interventions remain long-term in nature and remain prohibitively expensive for firms to acquire in the short-run.

feedback. This scenario is quite common in entrepreneurship, which is often characterized as a process of decision-making in the absence of data (Alvarez and Barney 2010). Without data, decision-makers increasingly rely on cognition to guide their search for risky and uncertain opportunities (Huang and Pearce 2015).[9] By focusing on how to creatively recombine existing choices (Baker and Nelson 2005, Sarasvathy 2001, Katila and Shane 2005) or to develop new alternatives altogether (Cromwell et al. 2018a, Huang and Pearce 2015), resource constraint may trigger broader, more interdependent changes. Together, I summarize the relationship between action resources and interdependent change in experimentation with the following hypotheses:

> **Hypothesis 1a**: As access to action resources increases, firms allocate these resources towards more experimental trials.
>
> **Hypothesis 1b**: As access to action resources increases, firms reduce interdependent changes per experimental trial.

2.3 Empirical Context: A/B Testing to Inform Interdependent Activities

I study differences between testing interdependent and independent changes in experimentation in the context of website A/B testing. Below, I detail describe the fundamentals of A/B testing and how it can be leveraged to test interdependent changes which help form strategy. I also describe the role of action resources in A/B testing.

2.3.1 Interdependent vs. Independent Changes in A/B Testing

In recent years, website A/B testing has emerged as a viable method for business to identify high-performing, strategic alternatives (Koning et al. 2019, Azevedo et al. 2019). In an A/B test, the experimenter sets up two conditions: "A," the control, is usually the existing website and "B," the treatment variant, represents a change (or set of changes) from the control that

[9] While often used interchangeably in extant research (Townsend et al. 2018), risk and uncertainty have distinct implications in entrepreneurship research. For reviews of extant perspectives, see e.g., Packard et al. (2017), Townsend et al. (2018).

tests a new alternative. Website visitors are randomly assigned to the two conditions, after which key metrics are computed and compared across the conditions. An example is provided in in Figure 2.1. Most commonly, the performance of the new alternative represented by the "B" variant is measured by an increase in the number of visitors who "convert" on an event. Online, the new alternative could involve new features, user interfaces (i.e., a new layout), and back-end changes (i.e., a new algorithm that offers user recommendations), among other examples (Kohavi and Thomke 2017).

Following Leiblein et al. (2018), a new alternative is strategic based on how interdependent that alternative is with the other activities that a firm is simultaneously pursuing. We can apply this definition in the context of website A/B testing to assess how strategic the alternative being tested is. Many online A/B tests involve simple changes, such as the color of a checkout button. Barring potential interdependencies with other visual elements (such as background colors), such changes are relatively independent in nature, insofar as the color of an individual icon has little bearing on the other activities that a firm is simultaneously pursuing. On the other hand, a more strategic alternative to evaluate in an A/B test involves a greater degree of interdependencies with other activities. For instance, should an online platform specializing in hotel bookings enter new verticals such as flights and car rental services in the travel industry (Thomke and Beyersdorfer 2018)? While such an experiment may involve several visual edits to a website's user interface, the most significant interdependencies triggered by this change involve those that are made to back-end website code (Kohavi et al. 2020). In the example of a hotel accommodation platform diversifying into new verticals, several simultaneous changes must be made to the websites' back-end data infrastructure to serve the user appropriate information on flights and car rental availability and pricing (Thomke and Beyersdorfer 2018). These changes require greater commitments of a company's scarce resources, such as its engineering talent to build and evaluate this alternative. Accordingly, by triggering more interdependencies, an A/B test may evaluate alternatives that are relatively more strategic in nature.

(a) Control Variant

(b) Treatment Variant

Figure 2.1: **A/B Testing and Opportunity Identification.** Maxis, the developer of SimCity, runs an A/B test to improve checkouts of the SimCity video game. Panel (a) displays the control variant, and Panel (b) displays the treatment variant. In this test, Maxis removes the promotional banner from the control variant, which is circled in red in Panel (a). The test results in a 43% increase in checkouts.

2.3.2 Action Resources in A/B Testing

Action resources in A/B testing can manifest themselves in many ways, such as data/analytics infrastructure, and a firm's access to testing software. Today, one of the most salient resource constraints limiting access to experimental action and running individual tests is website traffic (Azevedo et al. 2019). Resources for testing can be understood via a standard sales funnel. In conversion rate testing, customers begin their journey at the top of the funnel, where attention for a website's products and services is established. From here, the website continues to lose potential customers on the journey between the traffic that enters a website at the top of the funnel and where the sale or conversion is completed at the bottom of the funnel. While A/B testing can occur at any level of a website, my focus is on conversion rate testing, which seeks to improve the yield of customers who have already reached the bottom of the funnel and now must successfully convert in order for the firm to improve its performance. Here, performance in an experiment is measured in terms of *lift*, or the proportional increase in customers who successfully convert on a metric of interest, e.g., product purchases. Ultimately, the number of experimental trials that an organization can run at the bottom of the funnel will be constrained by the availability of traffic that enters the website at the top of the funnel.

2.4 Exploratory Analysis: Interdependent Change and Performance in Experimentation

In this paper, I develop and test theory for how interdependent changes are generated and evaluated via business experimentation. I begin with an exploratory analysis on the association of interdependent changes with experimental performance. The results of this analysis shall then motivate the second major empirical section in this paper on the effect of testing resources on interdependent change in experimentation, in Section 2.5.

2.4.1 Data

I obtain access to a novel, proprietary dataset of business experiments run using the widely-used, third-party A/B testing platform, Optimizely. While companies such as Google, Amazon, Microsoft and Booking.com built in-house experimentation platforms years ago, Optimizely helped pioneer easy-to-use A/B testing tools for technical and non-technical business professionals, enabling them design and implement their own business experiments.[10] Today, Optimizely provides A/B testing tools for more than one thousand clients, from start-up firms to Fortune 100 companies. As the market leader for A/B testing services, Optimizely provides access to online business experimentation for firms across industries (e.g., retail, finance, insurance, etc.), scale, and levels of technological sophistication.

I use data from the Optimizely X cloud platform, which provides conversion rate testing tools for web enterprises (see Section 2.3 for additional detail on sample tests and use cases). Optimizely archives data from A/B experiments run by firms on its cloud platform, including effect sizes, number of website visitors within an experiment, and experiment duration, among other factors. Of particular interest to my study is Optimizely's detailed microdata on changes made during an experiment. This change data enables us to create continuous measures for interdependent change in experimentation.

My unit of analysis in this exploratory analysis is the experiment: an A/B test. Despite the breadth and richness of data from Optimizely, it is an empirical challenge to distinguish true business experiments from other actions taken by firms on the platform. A basic requirement for a business experiment is that it generates information that helps the organization learn (Thomke 2003, p. 98) and that is relevant to a firm's strategy or configuration of choices (Rivkin 2000, Ghosh et al. 2020). Using criteria initially established in Ghosh et al. (2020), I classify true business experiments according to the following criteria: (1) at least one change per treatment variant (i.e., no A/A tests)[11], (2) at least one treatment

[10] While Optimizely offers services to help firms design their experiments, my dataset comes from the X platform where tests are designed, implemented, and analyzed independently by firms using the platform

[11] In an A/A test, the current practice is compared with itself. A/A tests are used to check the quality of

variant per test (i.e., no "hotfixes")[12], 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). An experiment ends during the last observed week in which scarce website traffic allocated to the experiment surpasses 1,000 visitors. To distinguish between business experiments and real options (Contigiani and Levinthal 2019), I examine the set of experiments that have successfully concluded and for which another experiment was confirmed to have followed the focal experiment within the year following my observation window.[13] Furthermore, I define outcomes at the experiment level 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 31,716 experiments run between April 2017 to March 2019 from Optimizely for our analysis.[14] Further detail on data construction can be found in Appendix A.1.

In addition, to augment this research and to triangulate available quantitative evidence on cognition and action in experimentation (Edmondson and McManus 2007, Jick 1979), I conducted semi-structured interviews with firms running A/B tests on Optimizely's X platform. Given varying perspectives on experimentation due to management seniority and experience (Ghosh et al. 2020, Leatherbee and Katila 2019), I conducted interviews with 21 A/B testing professionals across industries and at multiple hierarchical levels (e.g., developer, product manager, vice president, etc.). These interviews ranged in length from 45 to 90 minutes (average length: 54 minutes) and provided rich detail on decision making and resource allocation in A/B testing, among other factors.

---

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.

[12] 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.

[13] While it is theoretically important to distinguish real options from business experiments, my findings are robust to including "real option" tests in the sample, where a test with A and B variants runs indefinitely and without any follow-on tests.

[14] 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.

2.4.2 Variables

Dependent Variables

Given strategy formation's focus on identifying sets of activities that generate outstanding performance (Siggelkow 2002, Gavetti and Rivkin 2007), I follow the approach of papers on breakthrough innovation and operationalize top performance as performance within the top 5% of our sample. (Jung and Lee 2016, Kaplan and Vakili 2015, Singh and Fleming 2010).[15] Performance of an opportunity is measured in terms of lift, or the proportional improvement in conversion rates due to an experimental treatment. An example of a performance lift is the increase in sales of checkouts associated with a treatment (see Section 2.3 for examples). Lift offers a number of attractive features for measurement, such as the built-in normalization of effects which enables comparisons across experiments in terms of relative increases (Gordon et al. 2019). To capture a high-performance opportunity, I measure *Top 5% Lift*, which is a binary variable that takes a value of 1 should the lift be in the top 5% of the sample. Similarly, to measure performance failures, I measure *Bottom 5% Lift*, which is a binary variable that takes a value of 1 should the lift be in the bottom 5% of the sample.

Independent Variable

To capture complex, interdependent change in experimentation, I measure *Code Change*, which is the number of lines of customized JavaScript code added to implement a treatment tested in an A/B test. To appreciate why *Code Change* is an appropriate measure of interdependent change in experimentation, it is helpful to understand institutional context on how changes are typically implemented and tested on the Optimizely platform.

Optimizely's web experimentation platform offers a simple, "What You See Is What You Get" (WYSIWYG) development environment to design and build experimental treatments, where organizations can visually point and click on elements of their websites to

[15] While I follow the approach of prior literature in evaluating top performance according to the the top 5% of a sample, my results are robust to the use of other cutoffs, such as top 1% of lifts.

change.[16] For instance, a user may click on the background and change its color using the color toolbar. After recording their edits, the Optimizely platform will perform the corresponding back-end code changes to implement the treatment version of the website. This development environment is useful for simple treatments that do not involve interdependent changes.

On the other hand, treatments that test more complex, interdependent changes require back-end code customization on behalf of organizations. New alternatives invoking interdependent choices, such as a retailer's decision to enter a new market, often require substantial back-end code changes to a website in order to successfully implement (Thomke and Beyersdorfer 2018). For these cases, Optimizely enables developers to upload lines of JavaScript code to help implement these complex treatments. My interviews with A/B testing professionals suggest that the lines of custom code is a reliable measure for experiments testing interdependent decisions, requiring the scarce cognitive resource of engineers to appropriately code these treatments (see Table 2.1). As one UX designer reported, "Development resources are always hard to get, so I try to do whatever I can in the [WYSIWYG] editor...But we send the more complex builds to engineering, who would actually go through and code them up for us."

As the lines of JavaScript code increase, so too does the cognitive effort of engineers who must ensure that their complex code appropriately compiles without errors, enabling the experiment to run. An experiment that does not invoke any JavaScript code changes is measured with a *Code Change* value of zero. Additional detail on this measure with practical examples of how high-interdependence treatments drives the use of customized code is provided in Appendix A.1.

[16] An example of a "What You See is What Get" development environment is Microsoft Word, where text and content are laid out by the user in a way that resembles the final printed product. In contrast, LaTeX uses markup language to convert plain text input and commands into a final printed document.

Control Variables

I construct a number of variables that may associate with the interdependent change and performance of individual experiments. Following Ghosh et al. (2020), I control for an experiment's *Duration*, *Sample Size*, and *Variant Count*, each of which has potential to influence the performance of tests. In addition, to control for potential confounding influences due to multiple organizational performance objectives (Ethiraj and Levinthal 2009, Hu and Bettis 2018, Gaba and Greve 2019, Obloj and Sengul 2020), I measure *Metric Count*. Finally, to account for time invested in an experiment and possible diminishing marginal returns to successive testing (Ghosh et al. 2020), I control for *Development Time* and *Prior Experiments*. Measurement details, along with summary statistics and pairwise correlations for these variables are provided in Table 2.2.

2.4.3 Estimation Strategy

I analyze associations between complex, interdependent change and the performance of opportunities according to the following model:

$$Y_i = \beta(Code\ Change_i) + X_i B + \eta_i + \delta_i + \alpha_i + \epsilon_i.$$

Y_i represents the dependent variables of interest. η_i is an organization fixed effect that controls for time-invariant unobserved confounding factors (e.g., an organization's technical expertise, its industry, etc.), δ_i is a month fixed effect to control for potential shocks across all experiments run during the same time as the focal experiment (e.g., holiday shopping season), and α_i is a fixed effect for the primary metric type for experiment i. X_i is a vector of controls associated with an experiment i.

Our coefficient of interest is β, which estimates the effect of an increase in complex, interdependent change on Y_i. I estimate models using OLS with robust standard errors clustered at the organizational level, although my results are robust to the use of limited dependent variable models.[17]

[17] I 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).

Table 2.2: **Descriptive Statistics and Pairwise Correlations for Experiments** ($n = 31,716$).

Measurement of Variables

Variable	Measurement
Top 5% Lift	1 if lift in top 5%; 0 otherwise.
Bottom 5% Lift	1 if lift in bottom 5%; 0 otherwise.
Code Change‡	Count of lines of interactive JavaScript code; 0 if no lines written.
Duration	Number of weeks an experiment is active.
Experiment Traffic‡	Count of visitors included in the focal A/B test (in thousands.
Variant Count	Count of number of simultaneous treatment arms tested.
Metric Count	Number of metrics measured for the focal A/B test.
Development Time	Count of days from when experiment is initially created on the X platform and when it is run.
Prior Experiments	Number of experiments run the organization prior to the focal A/B test.

Summary Statistics and Pairwise Correlations

Variable	Mean	Std. Dev.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Top 5% Lift	0.05	0.22	0.00	1.00	1.00								
Bottom 5% Lift	0.05	0.22	0.00	1.00	-0.05	1.00							
Code Change‡	1.58	1.66	0.00	9.01	0.00	-0.05	1.00						
Duration	4.54	5.85	1.00	101.00	0.03	0.01	0.04	1.00					
Sample Size‡	8.61	1.37	6.91	15.95	-0.00	-0.04	0.09	-0.06	1.00				
Variant Count	2.43	0.93	2.00	23.00	0.06	-0.06	0.05	0.01	0.05	1.00			
Metric Count	5.71	6.76	1.00	84.00	0.02	-0.01	0.11	-0.05	0.02	0.03	1.00		
Development Time	11.38	24.03	0.00	240.00	0.01	-0.00	0.19	0.05	0.01	0.05	0.04	1.00	
Prior Experiments	30.21	54.19	0.00	591.00	-0.03	-0.03	0.05	-0.07	0.11	-0.05	0.04	-0.06	1.00

‡ Given rightward skew, I apply a natural log transformation to these variables.

2.4.4 Results

I begin by graphically examining the relationship between increased interdependent change in experimentation and the identification of high performance in an experiment in Figure 2.2, which displays a histogram of logged lifts that are color-coded according to the level of interdependent change of a given experiment.[18] High-interdependence experiments, which are experiments in the top 50% of *Code Change* are shaded in blue, while low-interdependence experiments, which are in the bottom 50% of *Code Change*, are shaded in red. Given my interest in observing behavior at the tails of the distribution which represent outstanding performance, I apply a natural log transformation to the y-axis of count of experiments. This helps us more easily compare and contrast behavior at the center of the distribution to the that of the right tail of high-performing lifts (where behavior in the tails would be hard to visualize in a regular histogram).

Figure 2.2 immediately offers two interesting descriptive results on nature of interdependent change and the performance of experiments across my sample. First, most experiments feature little interdependent change via *Code Change*, as the red bars that capture the bottom 50% of *Code Change* feature 3 lines or less of JavaScript code. Second, a plurality of experiments in the Optimizely sample feature lifts that are difficult to distinguish from zero (i.e., the center bar of the histogram), supporting the notion that most experiments fail to have an impact.

Further examining the figure, we find a compelling divergence in where low-interdependence experiments and high-interdependence experiments reside in the distribution of lifts. These differences between high-interdependence experiments (shaded in blue) and low-interdependence experiments (shaded in red) become apparent as we examine the distribution of lifts away

[18] Lifts in online A/B tests can be quite large. For instance, in a study of advertising measurement methodology, Gordon et al. (2019) report lifts from randomized control trials run on Facebook of three orders of magnitude in percentage points. For intuition, consider an experiment that increases conversion rates from 3% in control to 33% in the treatment. This ten-fold increase would represent a lift of $\frac{(.33-.03)}{.03} * 100 = 1,000\%$. Given the apparent right skew in lifts, it is instructive to examine logged lifts.

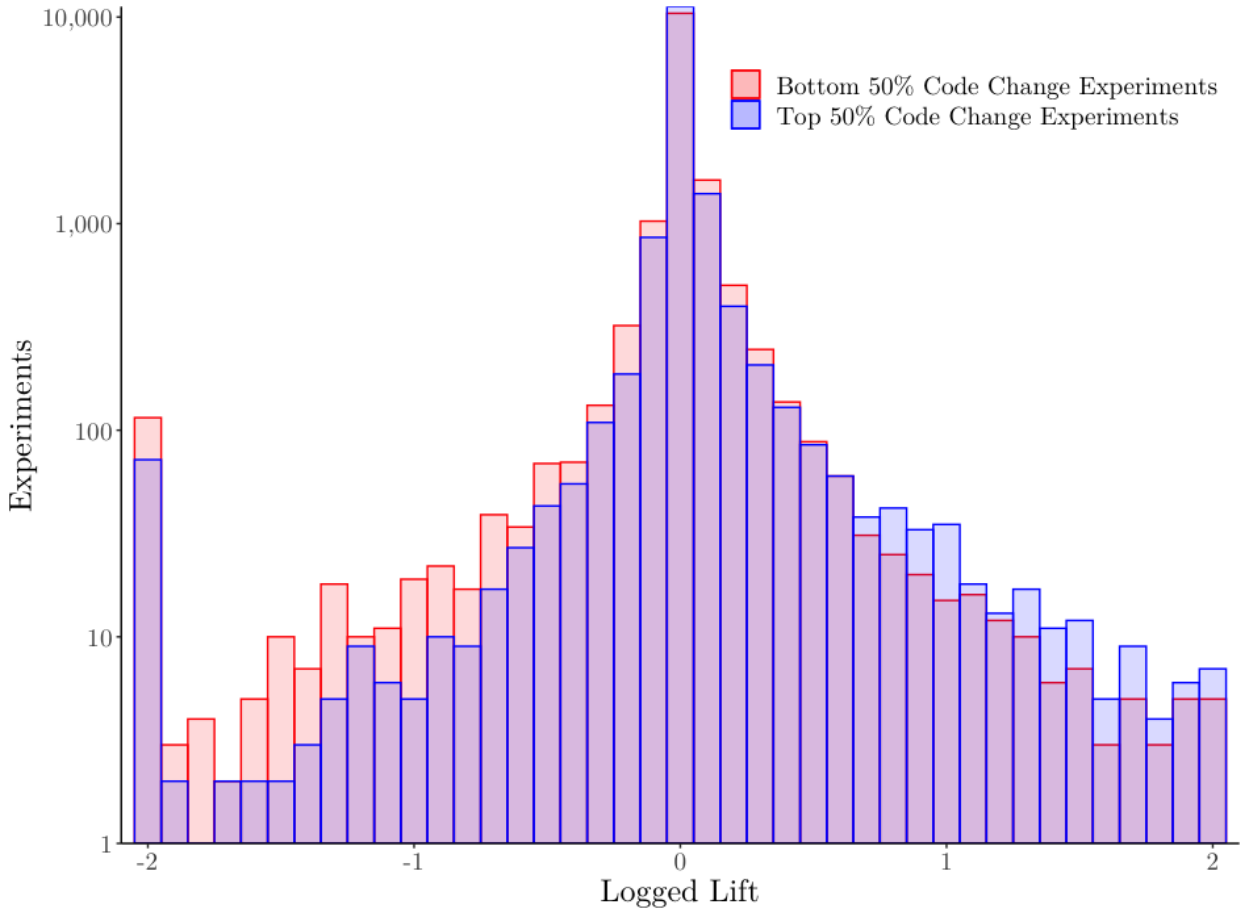


Figure 2.2: **Comparing Lifts from Low vs. High-Code Change Experiments.** Experiments that are in the bottom 50% of JavaScript code changes are displayed in red, while experiments that are in the top 50% of JavaScript code changes are in blue. Purple indicates an overlap between high- and low-code change experiments. A logarithmic transformation has been applied to the y-axis to highlight behavior at the tails of the distribution. Lift is displayed on the x-axis according to the transformation: $log(Lift + 1)$. The leftmost bar represents lifts of -0.99 or less, which are total losses in customer conversion as a result of an experiment.

from zero. A relatively larger proportion of low-interdependence experiments occupy the modest increases and decreases in lifts around zero, as evidenced by the prevalence of taller red bars. However, as we move further towards the right tail of high performance in the distribution, we see an increasing prevalence of high-interdependence experiments, represented by taller blue bars.

To further investigate the observed association between interdependent change in experimentation and large lifts, we turn to Table 2.3. The mean number of lines of custom JavaScript code for an experiment is 25. Thus, Model 2.3-1 demonstrates that a 100% increase in complex change, from 25 to 50 lines of custom code, is associated with a 0.5% increase in the chance of identifying a top 5% lift opportunity. Model 2.3-2 probes the robustness of this result with respect to experiment-level controls, and I find a similar effect of the complexity of change. To examine robustness to the 5% cutoff level, Models 2.3-3 and 2.3-4 examine associations between changes and the chance of a top 1% lift. Here, Models 2.3-3 and 2.3-4 illustrate an increase in the chance of a top 1% lift by over 0.2%.

While we find robustness for the association between the interdependence of change and top performance, is it possible that these experiments increase the chance of poor performance as well? As previously theorized by scholars of strategy and organization, the risks of tightly coupled, simultaneous changes include sharp performance failures (Levinthal 1997). If high-interdependence experiments were to naively increase technological uncertainty (Fleming 2001), for instance, we would expect a relatively symmetric increase in extreme lifts—both positive and negative.

Examining Figure 2.2, we find that high-interdependence experiments are *less* likely to populate the left tail of poorly performing experiments—as we find low-interdependence tests (in red) represent a greater proportion of experiments as lifts become increasingly negative. To formally test this observation, we turn to Table 2.4. Models 2.4-1 and 2.4-2 measure the association between experimental complexity and a lift that is in the bottom 5% of the sample. We find that increasing highly-complexity changes associates with a

Table 2.3: **Interdependent Change and Likelihood of Top Lift.** Ordinary least squares (OLS) estimation of experiment level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Top 5% Lift		Top 1% Lift	
	(1)	(2)	(3)	(4)
Code Change‡	0.00510****	0.00479****	0.00236****	0.00211****
	(0.00135)	(0.00133)	(0.00059)	(0.00057)
Duration		0.00046		0.00075****
		(0.00033)		(0.00022)
Sample Size‡		-0.00229*		0.00052
		(0.00121)		(0.00064)
Variant Count		0.00977****		0.00182*
		(0.00218)		(0.00100)
Metric Count		0.00029		0.00015
		(0.00032)		(0.00015)
Development Time		-0.00000		0.00003
		(0.00006)		(0.00003)
Prior Experiments		-0.00008**		-0.00005***
		(0.00004)		(0.00002)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Metric FE	Yes	Yes	Yes	Yes
Observations	31,716	31,716	31,716	31,716

Table 2.4: **Interdependent Change and Likelihood of Bottom Lift**. Ordinary least squares (OLS) estimation of experiment level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Bottom 5% Lift		Bottom 1% Lift	
	(1)	(2)	(3)	(4)
Code Change‡	-0.00279**	-0.00259**	-0.00200***	-0.00195***
	(0.00108)	(0.00111)	(0.00069)	(0.00073)
Duration		0.00014		0.00021
		(0.00037)		(0.00015)
Sample Size‡		-0.00252**		-0.00032
		(0.00103)		(0.00045)
Variant Count		-0.01397****		-0.00403****
		(0.00137)		(0.00067)
Metric Count		0.00004		0.00002
		(0.00030)		(0.00017)
Development Time		0.00006		0.00001
		(0.00007)		(0.00003)
Prior Experiments		-0.00003		-0.00001
		(0.00004)		(0.00003)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Metric FE	Yes	Yes	Yes	Yes
Observations	31,716	31,716	31,716	31,716

significantly decreased chance of a bottom 5% lift. Similarly, this association holds true in models 2.4-3 and 2.4-4 which demonstrate a decreased chance of a bottom 1% lift.

I run a number of additional analyses to probe the robustness of these results. First, while my hypothesizing and results demonstrate experimental complexity's positive association with performance breakthrough and negative association with performance failure, we may also be interested in complexity's effects on mean performance. Here, I test its association with mean lift and mean statistically significant lift and find consistent results—that is, complex, interdependent change associates with greater mean performance. Second, to assess the robustness of my measure for complexity, I construct an alternate measure for interdependent change in experimentation and find consistent results. Finally, while lift is the standard measure with which to measure performance in A/B testing, its weakness as

an outcome variable stems from its potential to overstate the absolute size of effects when faced with low baseline conversion rates. To address this concern, I subsample my data and reproduce my analyses for experiments without low baseline conversion rates and find consistent results. These tests are detailed in Appendix A.3.

2.4.5 Qualitative Evidence of Interdependent Change and Performance

When considering alternatives with uncertain returns, I find that increasing interdependent change in an experiment positively associates with the chance of identifying high-performance opportunities. Furthermore, increasing interdependent change in experimentation also associates with a decrease in the chance of performance failure. This is in contrast to a narrative of purely technological uncertainty due to change (cf. Fleming 2001), which would suggest that greater change in testing would increase the chance of both breakthrough and failure. Instead, I find that high-interdependence experiments bias clearly towards higher performance. Remarkably, despite these performance benefits, a plurality of experiments feature little to no interdependent changes in the form of code.

Managers are often acutely aware of the performance benefits of complex, interdependent change in experimentation, even while noting that these experiments do not occur frequently enough in their organizations. Noting the importance of cross-vertical experiments that combine a firm's interdependent, strategic choices, a principal product manager reported,

> "We define our big strategic initiatives, which then drive tests across different verticals that require a lot of engineering. . . These tests across different verticals—they're not that common—but we have a few, and they are very successful. . . Reflecting on this now does sort of probe a question that I want to take back to my team. And that's—why aren't we servicing cross-vertical opportunities as well?"

Given quantitative and qualitative evidence suggesting the performance benefits of interdependent change in experimentation, it is important to ask: what mechanisms may explain variation in interdependent change across experiments? Reflecting on this question, one A/B testing manager notes the central role of cognition in shaping high-performing,

interdependent changes:

> "You have to make this change relevant. You have to, you know, go big or go home—you have to have a bold change... Don't make this tentative little change that your users are not going to notice or respond to... And people sometimes take offense to that because I'm basically telling them 'your ideas are terrible.' I mean, their first ideas are *weak*... So I push them to really understand their user, and to really *think*. What other opportunities could they be testing instead? You know, how can I push them to think more and test higher-impact changes?"

In the next section, I explore how testing resources may influence the investment of cognition in experimental trails—thereby influencing the quality of interdependent change in experimentation.

2.5 Action Resources and Interdependent Change in Experimentation

In this section, I examine how action resources impact interdependent change in experimentation. I begin by presenting panel data analysis estimating the influence on an increase in action resources on interdependent change in experimentation. Then, to address potential endogeneity concerns, I identify and estimate the effects of a natural experiment which reduces action resources for a subset of organizations in my sample.

2.5.1 Data

I pair Optimizely's detailed data on A/B tests with that of SimilarWeb, a marketing intelligence platform which tracks detailed website traffic metrics for global websites. SimilarWeb provides detailed website performance metrics on web domains, including their pageviews and user engagement metrics (e.g., average visit duration of a user, "bounce rate" of users who navigate away from a page, etc.). While authors have used SimilarWeb data to proxy for performance associated with adopting A/B testing (Koning et al. 2019), I introduce a novel use of this data source to study website traffic as a resource of A/B testing. My use of this data source to represent action resources stems from the widespread practitioner understanding that website traffic is an input variable, rather than output variable, in conversion

rate testing. Of particular interest to my study is SimilarWeb's monthly data on pageviews and the source of traffic arriving to websites. I pull monthly data on these metrics from SimilarWeb's APIs.

With the panel of business experiments from the prior section, I merge in SimilarWeb data on monthly website traffic observed at the main domain or URL for a website. SimilarWeb records all traffic information associated with subdomains at the main domain level.[19] From Optimizely, I obtain the URL of the main domain associated with each experiment and perform an exact match on URLs. After performing this merge, I am left with 37,440 organization-month observations of organizational experimentation.

2.5.2 Variables

Dependent Variables

To capture how action resources impact interdependent change in experimentation, I measure both *outcomes* of action and cognition and experimentation and the antecedent cognitive *processes* that that occur prior to these outcomes.

Outcomes: Action and Cognition in Experimentation To measure how action resources influence organizational experimentation, I measure two pertinent outcomes describing the nature of experimentation in an organization across time. First, I measure *Experiment Count*, which is the number of experiments run by the organization in a given month. By counting the number of trials run per month, this is a measure of action in experimentation. Second, I measure *Mean Code Change*, which is the average lines of customized JavaScript code across all experiments run by an organization in a given month. This measures the investment of cognitive resources per experiment (see Section 2.4.2 for additional detail on how code changes demand scarce cognitive resources). Given the skewed nature of both of these outcomes, I apply a natural logarithmic transformation to each variable.

Process: Cognitive Evaluation To further examine how action resources may influence cognition, I propose two measures to capture cognitive search efforts prior to implementing an

[19] In other words, traffic to the URL mydomain.com/subdomain would be included in the traffic recorded for the URL mydomain.com.

experiment. The variation of alternatives prior to their evaluation plays an important role in cognition in experimentation (Csaszar and Laureiro-Martínez 2018, Gavetti and Menon 2016) (for additional detail, see Section 2.2.1). While internal variation and selection processes are difficult to observe in most empirical settings (Csaszar and Laureiro-Martínez 2018), I exploit the fact that many experiments are designed and considered on the Optimizely platform but that are ultimately not tested. Here, I measure two aspects of variation process. First, I measure *Max Code Change*, which is the maximum number of lines of customized JavaScript code among all experiments that are designed but not necessarily tested each month. Conceptually, this measures the greater variation from status quo that is considered by the organization, which is key to performance (Gavetti and Menon 2016). Second, I measure *Selectivity*, which is the fraction of all experimental alternatives generated that are ultimately not selected to be tested with customers. This captures the intensity with which are alternatives are evaluated and eliminated from consideration (Knudsen and Levinthal 2007, Keum and See 2017). Together, these measures capture the quality and quantity of alternatives generated prior to their testing.

Independent Variable

To capture the influence of action resources for experimentation, I measure *Traffic*, which is the logged number of monthly pageviews associated with a given web domain. Website traffic is a scarce resource for A/B testing which determines how many website visitors may be included in an experiment. A fundamental challenge in A/B testing is how firms choose to assign their total budget of visitors to different ideas to test via experiments (Azevedo et al. 2019). While traffic is evidently scarce for less-popular websites, even traffic-rich websites such as Google argue that its resource of visitors is scarce. For instance, large incumbents such as Google are often interested in precisely estimating small effect sizes, a testing strategy that requires very large sample sizes for each experiment (Azevedo et al. 2019, Deng et al. 2013). I observe website traffic at the main domain level, or at the start of the website conversion funnel, as described in Section 2.3.

Control Variables

I control for a number of factors that correlate with the quality of traffic received by organizations in my sample, in addition to a number of experimentation program characteristics. These factors may correlate with an organization's website traffic and experimentation outcomes, and therefore must be controlled in order to help produce reliable estimates of the effect of changes in action resources on interdependent change in experimentation. Regarding traffic characteristics that may influence cognition and action in experimentation, I control for *Pages Per Visit*, which is used by website administrators to understand their visitor engagement, which may influence website testing behavior.[20] In addition, I control for *Direct Leads*, which is the proportion of traffic that a website receives from non-referred sources. This helps control for baseline interest in a website (e.g., brand awareness) and its potential to enable or constrain web experimentation. Finally, to control for experimentation triggered by changes in visitor demographics, I control for *EU share*.[21]

To control for potential changes to a firm's cognitive resources for experimentation over time, I construct a number of time-variant program-level controls. For instance, I control for *Infrastructure Testing*, which measures how many testing infrastructure tests are run by an organization each month. This helps capture the maturity and quality of an organization's experimentation culture (Thomke 2020). As another measure for the sophistication of testing culture, I control for an organization's use of underpowered tests via *Underpowered Testing*. Finally, I control for the potential influences due to multiple organizational performance objectives and time invested in individual experiments via *Metric Count* and *Mean*

[20] Related measures include the duration of visit for the average user and bounce rate, which measure the rate at which web visitors immediately leave a website upon navigating to it. Given multicollinearity across each of these measures, I elect to use *Pages Per Visit* as a control, although my results are robust to the use of either duration or bounce rate in its place.

[21] When hosting visitors from other geographies and/or markets, website administrators often must change elements of their websites to accommodate differences due to language, culture, and other factors. Furthermore, web visitors from different jurisdictions may require additional privacy protections that trigger website changes, such as the protections mandated for EU citizens due to the General Data Privacy Regulation (Johnson et al. 2020), which occurred in the middle of my observation period. To control for a website's exposure to time-variant geographical changes in their audiences, I control for the proportion of website visitors from the European Union.

Development Time, respectively. These factors can influence the design of individual tests (Kohavi et al. 2020). To avoid the potential for "bad" control where potential outcomes are included as controls, I lag each of the aforementioned program-level controls (Angrist and Pischke 2008). Measurement details of these variables, including descriptive statistics and pairwise correlations, are displayed in Table 2.5.

2.5.3 Estimation Strategy

Using organization-month panel data, I estimate the effect of action resources on experimentation using the following model:

$$Y_{it} = \beta Traffic_{it} + \alpha_i + \delta_t + X_{it}B + \epsilon_{it}.$$

Y_{it} represents the dependent variables, such as *Code Change*. α_i is an organization effect that controls for time-invariant unobserved confounding factors (e.g., an organization's technical expertise, its industry characteristics, etc.), and δ_t is a month fixed effect to control for potential shocks across all organizations during the observation period (e.g., increased website traffic during the holiday season).

Our coefficient of interest is β, which estimates the effect of an increase in experimental action resources on Y_{it} within organizations over time. It is important to note that at time t, site-level traffic is assigned to an experimenting manager and cannot be endogenously influenced by experimenting managers in the short-run. Thus, β estimates the average treatment effect of increases in action resources on firm experimentation behavior. I estimate models using OLS with robust standard errors clustered at the organizational level.

Table 2.5: **Descriptive Statistics and Pairwise Correlations of Experimentation Programs** ($n = 37,440$)

Measurement of Variables

Variable	Measurement
Experiment Count‡	Monthly count of A/B tests run by the focal organization.
Mean Code Change‡	Mean count of lines of interactive JavaScript code; 0 if no lines written.
Max Code Change‡	Max count of lines of interactive JavaScript code written (but not necessarily tested) in a month; 0 if no lines written.
Selectivity	Percent experimental ideas that are considered but ultimately not approved to test.
Traffic‡	Monthly domain-level pageviews for an organization's website.
Pages Per Visit	Monthly average pages per visit for customers visiting the organization's website.
Direct Leads	Monthly proportion of site traffic originating from direct sources.
EU share	Monthly proportion of site visitors originating from the European Union.
Infrastructure Testing	Lagged monthly count of A/A tests run by the organization.
Underpowered Testing	Lagged monthly count tests run by the organization below 1,000 visitors.
Metric Count	Lagged monthly count of number of metrics measured by organizations across experiments.
Mean Development Time	Lagged monthly average development time of experiments.

Panel B: Organization-Month Level

Variable	Mean	Std. Dev.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment Count‡	0.48	0.73	0.00	3.91	1.00											
Mean Code Change‡	0.76	1.51	0.00	8.84	0.67	1.00										
Max Code Change‡	0.79	1.56	0.00	8.86	0.71	0.98	1.00									
Selectivity	0.66	0.36	0.00	1.00	-0.68	-0.49	-0.50	1.00								
Traffic‡	13.11	2.40	4.57	19.37	0.06	0.04	0.05	-0.01	1.00							
Pages Per Visit	6.27	5.37	1.00	136.35	0.02	0.01	0.01	-0.02	0.18	1.00						
Direct Leads	0.41	0.18	0.01	1.00	0.02	-0.02	-0.01	-0.03	0.02	0.23	1.00					
EU share	0.31	0.37	0.00	1.00	0.01	0.07	0.07	-0.02	-0.12	0.05	0.01	1.00				
Infrastructure Testing	0.21	1.47	0.00	80.00	0.18	0.10	0.12	-0.05	0.05	0.02	0.03	-0.01	1.00			
Underpowered Testing	0.82	3.41	0.00	74.00	0.16	0.15	0.15	-0.30	0.04	0.00	-0.01	0.03	0.04	1.00		
Metric Count	4.45	5.29	0.00	84.00	0.01	0.07	0.06	-0.03	-0.02	-0.01	-0.03	0.04	-0.02	0.00	1.00	
Mean Development Time	31.63	46.64	0.00	240.00	0.33	0.29	0.30	-0.26	0.03	-0.02	0.04	-0.00	0.12	0.19	-0.00	1.00

‡ Given rightward skew, I apply a natural log transformation to these variables.

2.5.4 Main Results

Having demonstrated the relationship between increasing complex, interdependent change and high performance in experiments in Section 2.4, we now examine how action resources in experimentation may influence interdependent change in experimentation. Table 2.6 displays associations between increasing action resources and experimental variation. Model 2.6-1 demonstrates that as traffic increases, organizations are more likely to increase experimental action via marginally more experiments each month. This intuitively suggests that as websites receive more pageviews (the resource of access to customers, at the very beginning of the conversion funnel), they have a bit more leeway for additional tests. This finding is robust to controls for website traffic quality and program characteristics, as displayed in Model 2.6-2.

Models 2.6-3 and 2.6-4 test our central question regarding the relationship between action resources and interdependent change. I find that increases in website traffic associate with significantly reduced interdependent change per experiment. In particular, it is interesting to note that the rate of decrease in interdependent change exceeds that of the increase in experiments per month.

Together, these results demonstrate the influence of action resources on decreasing interdependent change in experimentation. We now examine how resources may influence antecedent processes of cognitive evaluation prior to testing. Here, generating and evaluating alternatives prior to running a test plays a key role in cognition (Csaszar and Laureiro-Martínez 2018).

Table 2.7 displays the influence of resources on cognitive process. Model 2.7-1 measures the maximum complexity change of considered tests in a month. I find that as resources increase, organizations reduce the maximum code change of all ideas generated, thereby reducing a critical aspect of the quality of variation. Next, we examine how selective organizations are in their experimentation efforts in Model 2.7-3. I find that as resources for experimentation increase, organizations become less selective about which ideas they

Table 2.6: **Action Resources and Interdependent Change in Experimentation.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Experiment Count‡		Mean Code Change‡	
	(1)	(2)	(3)	(4)
Traffic‡	0.047**	0.048**	-0.097***	-0.093***
	(0.023)	(0.021)	(0.036)	(0.035)
Pages Per Visit		0.001		-0.004
		(0.002)		(0.004)
Direct Leads		0.030		-0.044
		(0.076)		(0.136)
EU share		-0.553****		-0.331
		(0.141)		(0.297)
Infrastructure Testing		0.030**		0.023*
		(0.014)		(0.013)
Underpowered Testing		0.018****		0.034****
		(0.004)		(0.007)
Metric Count		-0.003		-0.000
		(0.002)		(0.004)
Mean Development Time		0.003****		0.006****
		(0.000)		(0.001)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	37,440	37,440	37,440	37,440

ultimately choose to test—by rejecting fewer alternatives prior to testing.

Table 2.7: **Action Resources and Cognitive Process.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Max Code Change‡		Selectivity	
	(1)	(2)	(3)	(4)
Traffic‡	-0.095**	-0.090**	-0.031***	-0.025**
	(0.038)	(0.038)	(0.011)	(0.010)
Pages Per Visit		-0.005		-0.000
		(0.004)		(0.002)
Direct Leads		-0.082		-0.020
		(0.138)		(0.041)
EU share		-0.307		0.163**
		(0.310)		(0.072)
Infrastructure Testing		0.029**		-0.001
		(0.014)		(0.002)
Underpowered Testing		0.031****		-0.035****
		(0.007)		(0.003)
Metric Count		-0.001		-0.001
		(0.004)		(0.001)
Mean Development Time		0.007****		-0.001****
		(0.001)		(0.000)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	37,440	37,440	37,440	37,440

To probe the robustness of these findings, I conduct a series of stress tests. First, given that my primary dependent variables are count variables, I rerun my analyses using a conditional Poisson regression and find consistent results. Second, my interviews with A/B testing managers suggest some may respond more directly to testing resources in the form of the total unique visitors who visit their website, as opposed to total summed pageviews. I therefore construct another measure of traffic measuring unique visitors and find consistent results. Details and results from these tests are described in Appendix A.4.

In summary, as resources for experimentation increase, organizations run more experiments, but the relative complexity of these experiments drops at a greater rate. While

site-level traffic is a reasonably exogenous resource to an experimenting manager in the short-run (i.e., a manager does cannot directly set the level of this resource), in the long-run a firm's quality of experimentation may influence its level of website traffic. For instance, a firm's success in the market may influence the traffic it receives, and a successful firm may be less inclined to make complex, interdependent changes from existing practice. In addition, despite my efforts to control for cognitive resources such as the sophistication of organizational culture for experimentation, unobserved resources for experimental cognition may also vary over time, which may bias my results. To alleviate these concerns and probe the robustness of my findings, in the next section I describe and analyze a natural experiment which varies the resources made available for experimentation to organizations.

2.5.5 Robustness Check: Google Natural Experiment

To address endogeneity concerns about unobserved factors influencing firm website traffic and experimentation outcomes, from my interviews with experimentation managers, I have identified a natural experiment which quasi-exogenously reduced the traffic received by a subset of organizations in my Optimizely sample. In June 2018, Google implemented a video carousel on the search engine results page (SERP). This change shifted many results that were previously captured in organic blue links and placed them into a video carousel, as depicted in Figure A.2. Prior studies have demonstrated that the rearrangement of content on the SERP can have meaningful implications for website performance, as users may increase or decrease their propensity to click on a link in results based on its positioning (Edelman and Lai 2016). With the addition of the carousel, some organizations reported steep reductions in traffic from Google, as users would click on video links at a lower rate than their counterfactual classic organic link listings. From interviews and case studies published on the effects of the change, certain retail websites were particularly affected by the addition of the carousel, leading to steep reductions in referral traffic from Google (Gabe 2018). Further detail institutional detail on the Google video carousel change and its effects are detailed in Appendix A.5.

I use this natural experiment as a source of variation in action resources available for

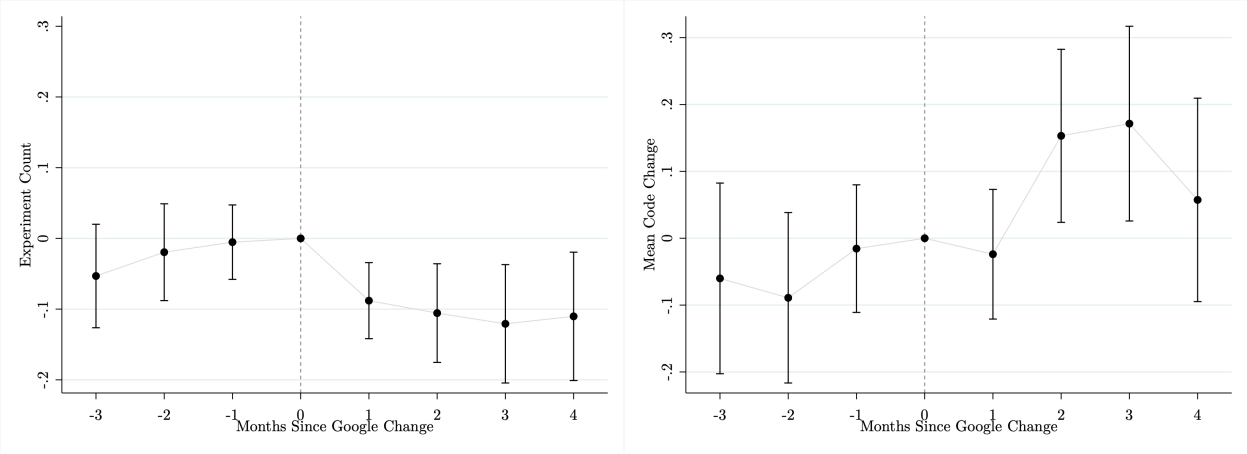


Figure 2.3: **Pre- and Post-Google Search Engine Results Change Estimates.** Plot of monthly estimates and 95% confidence intervals for the Retail difference-in-differences specification described in Section 5.5 (without controls). Subfigure (a) displays estimated effects on *Experiment Count*, while subfigure (b) displays effects on *Mean Code Change*.

experimentation. I estimate the effect of the change in two ways—first, in a difference-in-differences event study on the implementation of the SERP change on experimentation in retail, and second, via an instrument that estimates the effect of Google source traffic on overall website traffic.

Table 2.8 displays difference-in-differences estimates of the effect of the Google SERP change on retail organizations in the Optimizely sample. Models 2.8-1 and 2 demonstrate that a decrease in action resources associates with fewer experiments per month. In Models 2.8-3 and 4, we find that the shock decreasing action resources associates with a notable increase in mean code change. I investigate these trends graphically in Figure 2.3 which plots the estimates of the difference in outcomes in the months before and after the SERP change. Across both graphs, I find no significant differences in outcomes between retail and non-retail organizations. Furthermore, there is no evidence of pre-trends acting in the direction of the effect of treatment across measures of monthly experimentation and experimental code complexity.

While the difference-in-differences model give us increased confidence in the directional results of our main findings, it bears some shortcomings. Although retail firms were

Table 2.8: **Retail Difference-in-Differences Estimate of Google Natural Experiment.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Experiment Count‡		Mean Code Change‡	
	(1)	(2)	(3)	(4)
Retail × Post	-0.087***	-0.078**	0.131**	0.151**
	(0.033)	(0.033)	(0.065)	(0.063)
Pages Per Visit		0.001		-0.001
		(0.004)		(0.004)
Direct Leads		-0.124		0.090
		(0.122)		(0.190)
EU share		-0.026		0.304
		(0.171)		(0.310)
Infrastructure Testing		0.009		0.003
		(0.006)		(0.007)
Underpowered Testing		0.042****		0.065***
		(0.011)		(0.022)
Metric Count		-0.006		0.004
		(0.003)		(0.007)
Mean Development Time		0.002****		0.005****
		(0.000)		(0.001)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	12,480	12,480	12,480	12,480

more likely to be affected by the Google SERP change, are there unobserved, time-variant differences between firms in retail versus their counterparts not in retail that may potentially explain the difference in outcomes? Furthermore, our theoretical interest is in the effect of action resources on experimentation, which is not directly estimated with this approach.

To alleviate these concerns, I propose using an instrument in the form of Google source traffic to estimate the local average treatment effect of action resources on organizational experimentation. For Google source traffic to be a valid instrument, it must satisfy two conditions: a strong first stage and the exclusion restriction. A test of the strength of the first stage is displayed in Table 2.9. For the exclusion restriction, source traffic from Google must only affect experimentation through the channel of overall website traffic. My interviews demonstrate the Optimizely experimentation mangers overwhelmingly make experimentation decisions on the basis of total website traffic rather than the organic, unpaid traffic specifically coming from Google search. Furthermore, institutional factors such as the appearance of "cloaking" specifically penalize the use of Optimizely A/B testing in response to Google traffic, providing additional confidence on the exclusion restriction. This context is discussed in further detail in Appendix A.5.

Table 2.9 summarizes results from instrumental variable estimation of action resources and experimentation. Models 2.9-1 and 2.9-2 illustrate a strong first stage, where an increase in Google traffic strongly associates with an increase in overall traffic. The instrumental variable estimates of traffic and experiment count are presented in Models 2.9-3 and 2.9-4, with an estimated effect that is slightly larger than the corresponding estimate from the difference-in-differences estimation in Table 2.9 (note that the sign on the coefficient is flipped, where the difference-in-differences estimate corresponds to traffic decreases while the Google IV estimation measures *increases* in traffic). Similarly, Models 2.9-5 and 2.9-6 and displays an IV estimates of similar magnitude to the corresponding estimates from the difference-in-difference estimation in Table 2.8.

In summary, my estimates across methods demonstrate consistent findings—as re-

Table 2.9: **Google Traffic Instrument Estimate of Google Natural Experiment.** Instrumental variables (2SLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	First Stage		Second Stage			
	Traffic‡		Experiment Count‡		Mean Code Change‡	
	(1)	(2)	(3)	(4)	(5)	(6)
Google Traffic‡	0.528****	0.551****				
	(0.051)	(0.054)				
Traffic‡			0.106**	0.098**	-0.136**	-0.128**
			(0.053)	(0.049)	(0.067)	(0.064)
Pages Per Visit		0.015****		-0.001		0.002
		(0.003)		(0.004)		(0.004)
Direct Leads		0.758****		-0.106		0.059
		(0.106)		(0.119)		(0.191)
EU share		0.171		-0.069		0.361
		(0.131)		(0.171)		(0.318)
Infrastructure Testing		0.001		0.008		0.004
		(0.001)		(0.006)		(0.007)
Underpowered Testing		0.001		0.043****		0.063***
		(0.001)		(0.010)		(0.022)
Metric Count		-0.001		-0.005		0.004
		(0.001)		(0.003)		(0.007)
Mean Development Time		-0.000		0.002****		0.005****
		(0.000)		(0.000)		(0.001)
Organization FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,480	12,480	12,480	12,480	12,480	12,480
F-statistic	192.5	120.8				

sources for experimentation increase (decrease), organizations run marginally more (fewer) experiments, while reducing (increasing) the testing of interdependent changes per experiment.

2.5.6 Qualitative Evidence

In summary, I find that the resources available for experimentation play a driving role in experimental cognition and action. As resources increase, firms test marginally more experiments, but the mean complexity of these experiments suffer. In addition, I find evidence suggesting that as resources increase, organizations reduce cognitive, offline evaluation of alternatives prior to running an experiment. Measuring how organizations generate alternatives, I find that additional resources lead organizations to become less selective about testing alternatives and to limit the amount of change from the status quo.

In contrast, as the resources for experimentation fall, organizations run fewer but more complex tests. Given their role in screening for high-performing opportunities, this suggests a potential mechanism by which firms can broaden their search—by limiting access to action resources in experimentation. In particular, reduced action resources may prompt teams to engage with bolder, high-complexity alternatives first. A senior product manager shared her process as the following:

> "I keep a backlog of potential opportunities and continually reprioritize it... For the big, highly-prioritized opportunities, I work on the most impactful components first. If those turn out well, I accelerate the initiative. If they turn out badly, I don't scale, or I pivot, or I kill it... With less traffic, I focus on the high-impact opportunities first... And it's the thinking and discussion process [with my team] combined with scenario thinking that prevents big blindsiding in terms of missed opportunities."

The manager offers three critical insights. First, the manager notes that the organization faces a constant supply of opportunities to pursue. This makes the prioritization of which opportunities to test particularly important (Gans et al. 2019). Second, she goes on to note the role that action resources (traffic) play in prioritization, helping her team focus first on high-impact changes first. Third, and most importantly, she offers potential

mechanisms that help guide the prioritization of opportunities to pursue. By highlighting scenario thinking with her team, she recognizes the use of cognitive, offline methods to evaluate alternatives (Delmar and Shane 2003). She then suggests these cognitive efforts help her team avoid missing profitable opportunities.

2.6 Discussion and Conclusion

What is the role of testing interdependent changes in business experimentation? I address this question in two parts. First, I use proprietary data on website A/B tests to document how increased interdependent change in experimentation influences the identification of outstanding performance opportunities. Next, to explain variation in interdependent change in experimentation, I examine how the availability of feedback resources influences the trade-off between cognition and action in experimentation.

Below, I summarize my findings along with three major contributions of this work. I then conclude by highlighting limitations to my study while illustrating opportunities for future work.

2.6.1 Contributions

First, I contribute to a burgeoning literature on experimentation in strategy by highlighting an often-neglected reality of practice—business experimentation is resource-constrained. To contextualize my arguments, I introduce a conceptual framework for experimental resources that is broken down by the two phases of business experimentation—cognition and action. Using this framework, I focus on the sources of information for each experimental phase. Because feedback information is provided by customers, access to customers serves as the ultimate limiting condition to how many experiments an organization can run. This insight helps explain the anomaly of why some organizations feel experimentation remains costly (cf., Azevedo et al. 2019) in spite of the decreasing cost of testing infrastructure (cf., Koning et al. 2019). My conceptual framework also highlights how opportunity in experimentation may be realized in the cognition phase. This insight motivates the empirical analyses of my paper. In addition, this conceptual framework helps expose undertheorized aspects to

business experimentation practice. Later, I highlight a subset of these areas that I believe are promising for future study.

Second, I demonstrate how a previously neglected construct in business experimentation—interdependent change—can help unlock long-tail performance opportunities. Existing literature often focuses on how the frequency of experimentation, or action, may help screen for high-performing opportunities (Kerr et al. 2014, Azevedo et al. 2019). Nevertheless, this overlooks a crucial factor that differentiates the potential performance of each individual experiment—namely, the investment of scarce cognitive resources guiding the design of interdependent choices prior to testing. I find that as interdependent change in experimentation increases, organizations increase their chance of breakthrough while decreasing the chance of failure. My findings are consistent with narratives of cognition and strategic foresight (Csaszar and Laureiro-Martínez 2018, Gavetti and Levinthal 2000)—where managers consider the strength of alternatives and their potential interactions prior to testing them.

Third, my findings illustrate how scarce resources for experimentation help shape the trade-off between cognition and action in experimentation. As testing resources increase, firms run marginally more experiments, indicating increased action. Nonetheless, cognitive investments into these tests fall at a greater rate. Furthermore, as testing resources increase, organizations reduce the interdependent change of all alternatives generated while becoming less selective about which alternatives they choose to test. This has important implications for the formation of strategy, as strategy involves deciding what *not* to do (Porter 1996). With alternatives of reduced variation to compare up front, organizations may potentially hamper their ability to screen for high-performance opportunities (Levinthal 2017).

Nonetheless, my findings suggest a potential mechanism by which organizations may broaden their search for opportunities using experimentation—by limiting access to action resources. In particular, when faced with a shock that decreases action resources, organizations run experiments of greater interdependent change. My interviews suggest that fewer action resources help organizations prioritize their search effort by focusing on high-impact

opportunities. With fewer action resources, firms consider how to combine decisions to create bundles of choices that when combined together yield higher performance outcomes. Thus, limiting access to experimental feedback may help organizations appreciate the true opportunity cost of experimentation (Gans et al. 2019) by prioritizing what to test and what not to test. With limitations to testing, organizations may augment their rationality by considering more alternatives that differ from the status quo (Gavetti and Menon 2016). As we found earlier, these more complex, interdependent alternatives may help unlock long-tail performance.

2.6.2 Limitations and Future Work

I conclude by noting limitations to the present study and opportunities for future work. First, despite the strengths of my empirical setting in making experimental action and cognition observable, my findings are limited to enterprises with a web presence. An area of future work may be to study the generalizability of my findings to offline settings, such as nascent and resource-constrained markets (Baker and Nelson 2005, Katila and Shane 2005, McDonald and Eisenhardt 2019). Second, an implicit assumption to my analysis is that an organization's cognitive resources remain fixed in the short-run. I attempt to alleviate concerns of changing cognitive resources with a quasi-experiment that studies a shorter time period during which variation resources are plausibly stable. Nonetheless, relaxing this assumption may lead to fascinating future inquiries that study how cognitive resources such as organizational structure (Csaszar 2012), individual managerial cognitive ability (Helfat and Peteraf 2015), and an organization's experimentation culture (Thomke 2020) affect experimental performance. Finally, while the endemic complexity of the business environment is assumed to motivate the need for cognition in the search for high-performing opportunities (Gavetti and Levinthal 2000, Baumann et al. 2019), the particular nature of interaction patterns underlying the decision-making environment in A/B testing remains unobserved. Future work may examine how environmental complexity and unique interaction patterns in particular industries may shape how organizations should experiment to identify high-

performing opportunities.

Chapter 3

Do Senior Managers Help or Hurt Business Experiments? An Exploratory Study of Online Testing

Sourobh Ghosh, Stefan Thomke, Hazjier Pourkhalkhali

Abstract. Despite the adoption of business experiments to guide strategic decision-making, we lack a deeper understanding of how senior managers influence firm experimentation. Using proprietary data of experiments from a widely-used A/B testing platform, we explore how increasing seniority in management associates with learning and performance outcomes in experimentation. Our findings suggest that senior management's association is mixed. While senior managers' involvement associates with more complex experiments that create more significant learning signals, their involvement also associates with smaller performance improvements. Our results contribute to a burgeoning literature on experimentation in strategy, articulating associations of organizational design in data-driven decision-making and offering implications for how managers can modulate experimental search and performance outcomes.

3.1 Introduction

Strategy scholars have recognized the increasing use of business experiments to drive managerial decision-making (Andries et al. 2013, Camuffo et al. 2020a, Gans et al. 2019, Levinthal 2017, McDonald and Eisenhardt 2019). For example, the proliferation of A/B testing tools

has made experimentation an attractive method by which startups can test various elements of their value creation and capture processes (Koning et al. 2019). The benefits of an experimental approach include faster learning, improved performance, and reduced errors in strategic decision-making (Camuffo et al. 2020a, Thomke 2003). 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 senior management and organizational decision-making: If major strategic decisions are made with the help of experiments, then what is the job of senior managers? On the one hand, they may help organizations capture the learning and performance benefits of business experiments. Senior managers are well-positioned 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 pivots (Camuffo et al. 2020a), executive attention may help create and capture learning and performance benefits from experiments.

On the other hand, senior managers can bias decision-making and hamper these learning and performance benefits. For example, 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. Executives are also 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 et al. 2007). Accordingly, senior managers, through their positions of authority, may discourage the generation of new ideas (Keum and See 2017). Given the contradictory influences of senior management, we ask: how does increasing seniority in management associate with learning and performance

outcomes in experimentation?

To explore this question, we use a proprietary dataset of 6,375 business experiments run on the widely-used A/B testing platform, Optimizely. This global dataset of live business experiments includes start-ups, Fortune 500 companies, and public organizations across a wide range of industries. Descriptive in nature, this study offers representative cross-sectional analyses of business experimentation practice at scale and is among the first to document the influence of senior managers in business experimentation. 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 varying management seniority associates with different learning modes and performance outcomes. More senior management involvement associates 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 enables optimization and performance improvements.

This study offers a number of contributions to the literature on experimentation in strategy. First, we contribute to a burgeoning literature on experimentation in organizations (Camuffo et al. 2020a, Levinthal 2017) by linking senior management involvement to experimental performance outcomes. Second, our findings contribute to understanding potential limitations of organizational design in data-driven decision-making and search in the digital age (Brynjolfsson and McElheran 2016, Lee and Edmondson 2017, Puranam et al. 2014). Third, our empirical results describe different experimental learning modes in the formation of strategy, offering important implications for how managers can modulate organizational search and performance outcomes.

3.2 Experimentation, Strategy, and Structure

Organizations have a long 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 is now available to an entire organization and can be run in real-time, on live customers, to adjudicate firm-level commitments. This has profound implications as firms increasingly use experiments to test and form elements of strategy (Contigiani and Levinthal 2019, Leatherbee and Katila 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 et al. 2017). Emergent methods of strategy formation, such as bricolage and trial-and-error learning, prioritize signals from the external environment that determine the fitness of a strategy. While these approaches

may be effective at screening choices for their performance, the organization is often left with an incomplete understanding of *why* a certain set of activities yields superior performance, as a firm itself is often not the ultimate source of variation among activities. In contrast, deliberate strategy prioritizes intentionality, where the organization 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 fail. Thus, an experimental approach to strategy formation offers a middle ground between cognition and action (Levinthal 1997), where the organization 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 managers learn useful information about 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 et al. 1996, Murray and Tripsas 2004, Ries 2011). Experiments offer 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 through interim feedback, helps ventures pivot faster and avoids choices that yield false positive returns, or erroneous learning (Camuffo et al. 2020a). Experiments can also facilitate the identification of higher-performing choices (Gruber et al. 2008). In a study of A/B testing, a form of online experimentation, Koning et al. (2019) find that its adoption leads to increased product introductions and higher website performance over time.

3.2.1 The Effects of Senior Management Involvement

Management's influence on organizational decision-making can be exercised through many channels, such as increasing seniority (Bunderson et al. 2016, Joseph and Gaba 2020). To illustrate how management seniority can influence the testing of strategic decisions, consider the following example from one of the world's most visited travel accommodation platforms,

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. 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 organizational learning. But the impact of their influence is ambiguous: prior work has found that senior management can both benefit and impede organizational decision-making (Joseph and Gaba 2020).

Benefits of Senior Management

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 senior managers to address the resource allocation problem in strategy (Gavetti et al. 2012), enabling the controlled exploration of high-performing strategies. 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). Here, an executive can 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.

Given the influence of senior decision-makers on organizational initiatives (Gaba and Greve 2019, Joseph and Gaba 2020, Tuggle et al. 2010), when senior managers direct their attention towards and support an experimental approach to strategy, the firm may benefit from improved 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. 2020a). 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 Senior Management

In spite of the potential benefits of involving senior managers in experimentation, commonly cited risks 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 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 experiments to update them. 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 are overconfident in their beliefs, the organization may suffer from impaired learning and performance. In studying how information is passed up to management by subordinates within firms, Reitzig and Maciejovsky (2015) 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. This, in turn, can reduce the variation of ideas (Keum and See 2017). 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.

3.3 Method

Our study addresses the effects of senior management 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, for instance, 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 (Kohavi et al. 2020, Levinthal

2017).

A/B/n testing is a particularly useful setting to study the role of management seniority in strategic experimentation for several reasons. First, as seen in the Booking.com example, senior managers are often 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. 2020a, 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.

3.3.1 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, 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, 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 when they register with the Optimizely platform, enabling us to construct measures of a manager's position within an organization. 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. 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.[22]

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)[23], 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.[24]

3.3.2 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

[22] 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.

[23] 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.

[24] 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.

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

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.[25] 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. 2020a). 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.

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,

[25] See Appendix B for additional details on Optimizely's methodology for determining statistical significance.

including span of control (Rajan and Wulf 2006, Reitzig and Maciejovsky 2015), 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). Each of these measures captures different constructs, resulting in discrepancies in extant research on the influence of senior managers (Bunderson et al. 2016, Joseph and Gaba 2020).

To circumscribe our research, we specifically ask how experimentation may be influenced by managers from varying levels within their organizational hierarchies. Thus, we capture the effect of increasing steepness of hierarchy, which comes from 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 experimentation team. Users in an experimentation 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). 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. Because Specialist/Associates are present across all registered Optimizely experimentation teams, a higher Max Seniority score captures greater steepness between the highest-ranking individual(s) and the Specialist/Associates associated with an experimentation team.[26]

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

[26] 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.

At the organization level, we control for *Organization Age* through years since founding and Employee Count through the number of employees, both of which have been associated with organizational 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 organization, 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 3.1.

Table 3.1: **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	0.11	0.31	0	1	1										
Max Lift	0.10	0.67	-0.92	12.91	0.20	1									
Max Seniority	4.54	1.16	1	6	0.04	-0.03	1								
Variant Count	2.43	0.84	2	8	0.03	0.08	-0.06	1							
Max Variant Complexity	1.36	0.59	1	4	0.03	0.016	0.029	0.02	1						
Mean Variant Complexity	1.32	0.55	1.00	4.00	0.03	0.007	0.03	-0.08	0.97	1					
Duration	4.04	3.85	1	31	0.06	0.02	0.03	0.01	0.02	0.02	1				
Traffic	35.37	383.49	1.00	24,054.73	0.034	0.037	0.016	0.01	0.00	0.00	0.03	1			
Organization Age	33.24	37.57	1	282	0.05	0.02	0	0.02	-0.00	-0.00	0.02	0.01	1		
Employee Count	18,341.17	59,144.17	5	377,757	0.01	0.02	0.10	-0.05	-0.01	-0.01	0.06	0.05	0.35	1	
Technological Integrations	21.93	4.30	0	27	0.02	0.01	0.07	-0.04	-0.05	-0.05	0.01	0.02	-0.03	-0.06	1

3.3.3 Model Specification

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

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

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 increasing management seniority and experimental outcomes. We estimate models using ordinary least squares with robust standard errors clustered at the team level.

3.4 Results

Table 3.2 reports associations between increasing management seniority and outcomes. Model 3.2-1 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.[27] However, Model 3.2-2 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").

The results from Table 3.2 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 power increases with larger studied effect sizes. Nonetheless, it is possible that senior management's countervailing associations with lift and positive statsig may be the result of other mechanisms, which we explore in the following section.

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

Table 3.2: **Management Seniority 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)	Positive Statsig
	(1)	(2)
Max Seniority	-0.009**	0.01**
	(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]
Organization 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

3.4.1 Management Seniority and Design Choices

Design choices in experimentation can have a meaningful impact on learning modes and performance outcomes (Loch et al. 2001, Sommer and Loch 2004, Thomke et al. 1998). Is the mixed effect of senior management 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 et al. 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. 2020a), a heuristic that supports learning via incremental changes. Moreover, testing complex, tightly-coupled ideas may generate performance failures for the organization (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

Organizations 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 et al. 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.[28] 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 the chance of inference error due to increasing false discovery rates that arise in multiple hypothesis testing (Pekelis et al. 2015).[29] Besides the threats of reduced power, parallel testing may also impede strategic commitment by prompting organizations 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

[28] 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.

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

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 management seniority 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.[30] 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.

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)

3.4.2 Results: Management Seniority and Design Choices

To understand the mixed effect of management seniority on performance outcomes, we examine associations of increasing management seniority with design choices in Table 3.3. Model 3.3-1 tests the association between management seniority and *Max Variant Complexity*. In particular, we find that each increase of the hierarchical rank of the most senior person on an

[30] 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.

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 3.3-2 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 3.3-3, we test the association between management seniority and *Mean Variant Complexity*, which measures the average number of simultaneously tested elements across variants in an experiment. Similar to Model 3.3-1, we find that an increase in senior rank associates with an increase variant complexity, or 0.018 ($p = .018$) more mean changes per experiment.

To understand how the aforementioned design choices may influence learning modes and performance outcomes, we turn to Table 3.4. Model 3.4-1 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.[31] Model 3.4-2 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 variant 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 3.4-3 to 3.4-5 illustrate the association between design choices and positive statsig signals. Model 3.4-3 demonstrates that increasing the number of treatment variants has an association with *Positive Statsig* that is difficult to distinguish from zero ($p = .198$). Models 3.4-4 and 3.4-5 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

[31] 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.

Table 3.3: **Management Seniority 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
	(1)	(2)	(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]
Organization 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 3.4: **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) (1)	ln(Max Lift + 1) (2)	Positive Statsig (3)	Positive Statsig (4)	Positive Statsig (5)
Variant Count	0.038***		0.009		
	(0.007)		(0.007)		
	[0.00000]		[0.198]		
Max Variant		0.014*		0.019**	
Complexity		(0.008)		-0.009	
		[0.079]		[0.041]	
Mean Variant					0.021**
Complexity					-0.01
					[0.032]
Duration	0.001	0.001	0.008***	0.008***	0.008***
	(0.001)	(0.001)	(0.002)	-0.002	-0.002
	[0.409]	[0.358]	[0.00001]	[0.00002]	[0.00002]
Traffic	0.00001	0.00001	0.00003**	0.00003**	0.00003**
	(0.00001)	(0.00001)	(0.00001)	-0.00001	-0.00001
	[0.148]	[0.122]	[0.024]	[0.023]	[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

($p = 0.032$) in the chance of a *Positive Statsig* learning signal, respectively.

Taken together, these results demonstrate the senior management involvement has a mixed effect on learning modes and performance. Increased seniority 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, management seniority may inhibit incremental, parallel cause-and-effect learning.

3.4.3 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, organizations 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 team be dependent observations?

To address this question, we aggregate our analyses of management seniority at the level of the experimentation team and show the results in Table B.1. Model B.1-1 demonstrates that increasing seniority is associated a 4.6% decrease ($p = 0.039$) in conversion rates, while Model B.1-2 shows a 4.7% increase ($p < 0.001$) in the chance of Positive Statsig across experiments. Regarding design choices, Models B.1-3 and B.1-4 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 team 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 B.2, 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 management seniority'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 (Kohavi et al. 2020). 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 et al. 1999), or even the fact that websites become increasingly optimized over time and yield fewer opportunities for improvement. In Table B.3, we re-run our analyses from Table 3.2 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 3.2-1 to -0.7% in Model B.2-1), 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 3.2, demonstrating hierarchy's positive association

with positive statsig learning outcomes and negative association with lift.

3.5 Discussion and Conclusion

The proliferation of business experimentation in strategy formation has emerged as an important area of study (Azevedo et al. 2019, Camuffo et al. 2020a, Koning et al. 2019, Levinthal 2017, Thomke 2020); nonetheless, our understanding of how organizational structure may influence experimentation remains limited. In this paper, we explore how increasing management seniority within an experimentation organization influences organizational learning and performance outcomes. To study this question, we construct and analyze a novel, proprietary dataset of 6,375 experiments on the A/B/n testing platform Optimizely. We find that increasing management seniority in an experimentation team associates with more significant 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 nor curse to experimentation.

3.5.1 Organizational Structure and Experimentation

Following a rich literary tradition mapping the relationship between structure and performance in strategy (Csaszar 2012, Cyert and March 1963), we contribute to a burgeoning literature on experimentation in strategy (Camuffo et al. 2020a, Koning et al. 2019, Levinthal 2017) by exploring the relationship of organizational structure and experimentation outcomes. Our findings suggest a contradiction in senior management's influence on experimentation: whereas senior managers associate with discovery-driven learning and statistically significant learning signals, they negatively associate with high-performing experiments. 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 design choices and their payoffs (Levinthal 1997, Thomke et al. 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. 2020a) and ensure that they are climbing an appropriate hill. However, the risk of committing simultaneous changes (i.e., a long jump) is a large loss in performance 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, p. 284) 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 et al. 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. Recent empirical evidence suggests that managers who oversee several projects (such as those further up a hierarchy) accept greater risks in search. They understand that while any individual project may have a higher chance of failure, a portfolio of riskier projects may yield greater long-run high performance (Eggers and Kaul 2018, Joseph et al. 2016). In contrast, junior-driven organizations may have near-term incentives in experimentation, influencing their search behavior (Lee and Meyer-Doyle 2017). Measuring the effect of management seniority and design choices on long-term performance outcomes is an important area for future study.

3.5.2 Management Seniority and Organizational Search

Our findings also contribute to understanding the limitations that organizational design may place on decision-making and search in the digital age (Brynjolfsson and McElheran 2016, Lee and Edmondson 2017, Puranam et al. 2014). Inexpensive business experimentation has captured the attention of academics and practitioners alike, promising a data-driven approach to help organizations overcome their tendency to consider too few alternatives in search (Levinthal 2017, Ries 2011). Despite this promise, our findings align with prior work which finds an overall chilling effect on the variation of alternatives considered by senior managers, where organizations with increasing hierarchy or steepness of authority consider fewer projects and ideas (Csaszar 2012, Reitzig and Maciejovsky 2015). In our setting, increasing the rank of the most senior member of an experimentation organization associates with fewer variants per experiment—hence reducing the number of simultaneously considered alternatives.

While our data is particularly well-suited to observing a structure's effect on variation, we cannot observe its influence on the process of selection (Keum and See 2017). In particular, do senior managers impede the selection and retention of high-performing alternatives? Future work should examine whether potential biases from the "HiPPO" effect influence an experimenting organization's ability to adapt to a changing external environment over time.

Furthermore, understanding how alternate mechanisms may moderate the relationship between management seniority and variation may suggest solutions for the HiPPO effect. For instance, can senior managers delegate decision-making to junior members of an organization who may have specialized insight on new opportunities (Dobrajska et al. 2015)?

3.5.3 Learning Modes in Experimentation

Finally, 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 suggest 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 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.

3.5.4 Limitations and Future Work

We conclude by noting limitations to the present study and additional opportunities for future work. First, our study offers descriptive findings 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, we must caution that are findings are not causal and 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 may differ. Second, while our data lends itself well to conceptualizations of the steepness of hierarchy, it does not directly capture other potential constructs of interest, such as cross-relationships (Bur-

ton and Obel 2004), span of control (Rajan and Wulf 2006), or tallness (Lee 2020). Here, follow-on research could examine the influence of these alternate mechanisms. Finally, our findings on senior management'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.

Chapter 4

Iterative Coordination and Innovation: Prioritizing Value over Novelty

Sourobh Ghosh, Andy Wu

Abstract. An innovating organization faces the challenge of how to prioritize distinct goals of novelty and value, both of which underlie innovation. Popular practitioner frameworks like Agile management suggest that organizations can adopt an iterative approach of frequent meetings to prioritize between these goals, a practice we refer to as *iterative coordination*. Despite iterative coordination's widespread use in innovation management, its effects on novelty and value in innovation remain unknown. With the information technology firm Google, we embed a field experiment within a hackathon software development competition to identify the effect of iterative coordination on innovation. We find that iterative coordination causes firms to *implicitly* prioritize value in innovation: while iteratively coordinating firms develop more valuable products, these products are simultaneously less novel. Furthermore, by tracking software code, we find that iteratively coordinating firms favor integration at the cost of knowledge-creating specialization. A follow-on laboratory study documents that increasing the frequency and opportunities to reprioritize goals in iterative coordination meetings reinforces value and integration, while reducing novelty and specialization. This chapter offers three key contributions: highlighting how processes to prioritize among multiple performance goals may implicitly favor certain outcomes; introducing a new empirical methodology of software code version-tracking for measur-

ing innovation process; and leveraging the emergent phenomenon of hackathons to study new methods of organizing.

4.1 Introduction

Organizations often face the challenge of simultaneously pursuing multiple performance goals (Cyert and March 1963, Gavetti et al. 2012). For instance, airlines simultaneously strive for safety and profitability (Gaba and Greve 2019), while manufacturing firms seek to concurrently decrease costs and increase revenues (Obloj and Sengul 2020). Often, progress made in pursuit of one goal may inadvertently undermine performance towards other goals (Hu and Bettis 2018). This challenge applies broadly to organizations—even where individuals do not have conflicting preferences per se (Ethiraj and Levinthal 2009)—because an organization's multiple goals do not perfectly correlate with one another (Simon 1964). To help manage the pursuit of multiple goals, organizations can prioritize their most important subset of goals first before addressing those of lesser importance (Ethiraj and Levinthal 2009, Unsworth et al. 2014). Nonetheless, how organizations manage the pursuit of multiple goals for which there is no clear ex ante prioritization available remains unclear in existing organizational research (Greve and Gaba 2020).

One situation where organizations must manage multiple simultaneous goals with no clear, established prioritization among them is the pursuit of innovation. Scholars across literatures conceptualize innovation as the simultaneous pursuit of novelty and value (Amabile 1983, Kaplan and Vakili 2015, Singh and Fleming 2010). Despite novelty and value being distinct dimensions of innovation performance, prior work simplifies innovation to a singular dimension by implicitly assuming that the two dimensions of novelty and value travel with one another (Oldham and Cummings 1996, Shalley and Perry-Smith 2001). Nonetheless, more recent literature suggests that novelty and value may diverge in practice (Berg 2014). For instance, a mobile application that translates words from an alien language to English may be novel but may offer little value, whereas a simple mobile payments application may be quite valuable to customers, albeit not entirely novel. Thus, we take the view that the

process of innovation may be better-conceptualized as the pursuit of the distinct goals of novelty and value. Yet because an innovating organization must achieve both novelty and value, it is difficult to ex ante prioritize between the two.[32] For instance, for a firm striving to develop an innovative mobile application, does the firm develop an application of high value that would be of known interest to customers leading to its purchase, or does it instead focus on novelty to help differentiate their application in a crowded market? In short, it is unclear how an organization would prioritize between these two distinct goals. This challenge raises the question of which techniques an organization may use to prioritize among the underlying distinct goals of novelty and value when striving for innovation as an outcome.

Practitioners of popular management frameworks such as Agile management prescribe an iterative approach of frequent meetings to prioritize among multiple goals in innovation (Sutherland and Sutherland 2014, Rigby et al. 2016a, Bernstein et al. 2019), a practice we refer to as *iterative coordination*. In contrast to traditional innovation management practices which emphasize extensive ex ante planning to manage multiple competing objectives, Agile broadly encompasses a set of management practices defined by an "iterative approach [which] makes it easier to keep projects aligned." (Relihan 2018). As part of this iterative approach, Agile practitioners and gurus prescribe the use of iterative coordination in the form of frequent "stand-up" meetings for prioritizing among multiple goals in innovation (Strode et al. 2012, Rigby et al. 2016a, Sutherland and Sutherland 2014, pp. 79-80).[33] Helping drive its widespread use in practice for the management of innovation, practitioners often adopt iterative coordination with the expectation that it will help their firms ultimately produce more

[32] Following convention in the literature (e.g., Amabile 1983, Kaplan and Vakili 2015), we define innovation as something that is both valuable and novel. Under this definition, innovation emerges as a product of the pursuit of both novelty and value. Novelty for its own sake—such as a mobile application that translates words from an alien language to English—offers little value that can be created and captured by the organization. An innovative idea must involve the potential for an organization to commercialize it, as has been long assumed prior literature (e.g., Kaplan and Vakili 2015). On the other hand, a simple mobile payments application that is valuable but not novel does not qualify as innovation per the definition we follow.

[33] A stand-up meeting allows organizational members to discuss organizational priorities and tasks in a short meeting, often while standing up. Figure 4.1 situates stand-up meetings within a family of related practices.

innovation (Rigby et al. 2016a, Birkinshaw 2018, Bernstein et al. 2019). Nonetheless, the scholarly literature lacks both theoretical grounding and empirical evidence for how iterative coordination actually impacts innovation.[34] Motivated by the managerial importance of this practice and the striking absence of rigorous empirical evidence grounded in organizational theory to validate it, we ask: how does iterative coordination to manage innovation affect the outcomes of novelty and value?

In this paper, we develop theory for how iterative coordination affects an organization's ability to deliver novelty and value when innovating. While basic intuition would suggest that an organization could choose to pursue either novelty or value (or both), we argue that practicing iterative coordination drives processes that ultimately result in the prioritization of value over novelty in what is eventually delivered. Relative to a baseline of minimal coordination (Lifshitz-Assaf et al. 2020), as an organization uses additional meetings to discuss its goals, it creates additional interim deadlines (Gersick 1988, 1989, Waller et al. 2002). This leads individuals to focus on integrating their existing knowledge to create value as opposed to specializing in their work to create novelty. Finally, to avoid failure in the pursuit of their originally stated goals, the organization endogenously shifts its goals with each meeting to achieve value over novelty. Thus, although iterative coordination may appear to be an impartial way to manage goals while in the pursuit of innovation, we posit that the practice *implicitly* shifts an organization to realize value at the cost of novelty. The prioritization of value over novelty is implicit, since iterative coordination does not feature any explicit cues to favor value in goals and outcomes.

To empirically measure the effects of iterative coordination on innovation, we part-

[34] Recent organizations literature loosely relating to the construct of iterative coordination includes Obloj and Sengul (2020)'s study of managing multiple performance objectives in the context of the French manufacturing sector. Here, the authors find that frequent face-to-face meetings among executives help them manage their multiple goals, helping address trade-offs between their goals and iterate towards outcomes that are acceptable to all parties involved. Similarly, in analyzing results from their study of interdependent task environments in the automobile industry, Hu and Bettis (2018) argue in favor of using iterative design approaches to address unforeseen interdependencies across objectives. Neither study fully characterizes the construct of iterative coordination that we formally introduce in Section 4.2, and more importantly, they offer no predictions for how frequent meetings to prioritize among multiple goals may shape innovation goals and outcomes.

nered with Google LLC, a multinational information technology firm, to embed a field experiment within a public, one-day software application development competition, popularly known as a hackathon. We randomly assign firms at the hackathon to a treatment of iterative coordination. This exogenous variation mitigates traditional endogeneity concerns associated with archival data approaches (Chatterji et al. 2016). To collect precise data tracking firms, we introduce a novel methodology leveraging the version-control systems used in software development. By documenting the progress of actual software code developed, we capture patterns of firm activities at a granular level over time (by minute) in a balanced panel dataset. Our partner provided performance assessments of each organization's final software applications. In addition, we run a follow-up experiment in the laboratory where we vary mechanisms for iterative coordination and further validate the internal consistency of our findings.[35]

We find that, although iterative coordination leads firms to develop products that are judged to be more valuable, these products are simultaneously less novel. Furthermore, by tracking minute-by-minute changes in the software source code, we find that iteratively coordinating firms favor knowledge integration at the cost of in-depth, specialized knowledge creation by their members. In the follow-on laboratory study, we find that increasing the meeting frequency and opportunities for goal reprioritization in the implementation of iterative coordination reinforces integration and value, while reducing specialization and novelty.

Our study offers three contributions to the organizations literature. First, our findings contribute to the literature on managing multiple performance goals in strategy (Ethiraj and Levinthal 2009, Hu and Bettis 2018, Gaba and Greve 2019, Obloj and Sengul 2020) by studying how iterative approaches may direct organizations to prioritize among distinct goals in innovation. In particular, we emphasize the idea that iterative coordination, and iterative management practices more generally, might carry with it an implicit and overlooked consequence for which dimension of innovation is ultimately prioritized. Second, we con-

[35] We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

tribute novel methodology in the empirical study of organizations by introducing software code tracking as a new data collection method for studying organizational innovation. This methodology allows researchers to track innovation process at a granular level in real time—e.g., exact identification of occurrences of knowledge integration, etc.—bearing implications for the study of innovation across individuals, firms, and time. Finally, we contribute to an emergent literature on new forms of organizing (Levine and Prietula 2014, Puranam et al. 2014, Burton et al. 2017) by studying the management of innovation in hackathons, a setting that encapsulates the broader challenges facing organizations pursuing multiple, potentially conflicting objectives in innovation.

The structure of this chapter reflects the exploratory orientation we take toward our study of iterative coordination and innovation. Section 4.2 details the widespread managerial phenomenon and builds a general theory relating iterative coordination to potential mechanisms that might affect novelty and value in innovation. Our primary experiment in Section 4.3 then measures the general effects of iterative coordination on innovation in an externally valid field context. Based on the theoretical intuition that emerges from these empirical findings, Section 4.4 presents the follow-on laboratory experiment that validates specific mechanisms in a more controlled and precise environment. Finally, Section 4.5 details the contributions that emerge over the course of these theoretical development and experimental efforts.

4.2 Iterative Coordination and Innovation

4.2.1 Iterative Coordination Phenomenon in Practice

We propose a model of iterative coordination based on its implementation in practice. Iterative coordination is commonly practiced as a part of Agile management, a broad set of management practices defined by an iterative approach for prioritizing among multiple performance goals (Schwaber and Sutherland 2013, Rigby et al. 2016a).[36] Agile's key in-

[36] The use of the name "Agile" to describe this family of practices officially dates to February 2001, when 17 software developers gathered in Snowbird, Utah to develop the principles behind what would become the Agile Manifesto (Beck et al. 2001). This work inspired a family of alternative interpretations and

sight is the use of an iterative approach to prioritizing among multiple organizational goals (Rigby et al. 2016a, Relihan 2018). Seeking to rebut the extensive ex ante planning and linear approach of traditional manager-driven coordination methods, the designers of Agile methodology reasoned that more frequent iteration on organizational goals was necessary to help organizations embrace innovative outcomes (Cao and Ramesh 2007, Furr et al. 2016). To this end, existing methods for organizational coordination, such as traditional planning and role assignment by managers (Van de Ven et al. 1976, Okhuysen and Bechky 2009), were deemed antiquated and ineffective (Hoppmann et al. 2019, Lu et al. 2019). With new methods to rapidly iterate on organizational goals, Agile methodology could make organizations more flexible and adaptive to their environments, helping them innovate towards product-market fit (Ries 2011) and market differentiation to achieve competitive advantage (Vesey 1991). While Agile prescribes a variety of practices to help organizations iterate on multiple performance goals, the use of iterative coordination remains a unifying factor across these diverse implementations. To illustrate the distinction between our specific notion of iterative coordination and Agile more generally, Figure 4.1 summarizes the broader set of iterative practices most commonly associated with Agile, of which iterative coordination represents an important subset.

derivative frameworks that sometimes are referred to as Agile despite not exactly matching the original Manifesto. Our use of the term Agile refers to the common aspects of these varied interpretations, and not specifically to the Agile Manifesto.

Agile Practice	Description
Interim Deadline Frequency	
Short Iterations*	Breaking the development process down to phases, each with an interim deadline.
Frequent Releases	Setting nearly continuous deadlines for delivery of working product features to the customer.
Prioritization Discussion Content	
Pre-Existing Goal	
Iteration/Sprint Review*	Reporting of completed work and its efficacy towards achieving pre-defined performance objectives.
Retrospective	Allowing opportunity to reflect on which general processes were ineffective towards a pre-existing goal and can be improved.
Interim Goal	
Iteration/Sprint Planning*	Determining which tasks must be completed to address short-term goals within the context of a development cycle, known as an iteration or sprint.
Team Estimation	Estimating, formally or informally, the effort or cost associated with a potential development goal.
Release Planning	Identifying which aspects of a product or service in question will be delivered to customers for feedback.
Long-Term Goal	
Product Roadmapping*	Defining long-term goal to overlay onto milestones for how the product or service will be developed (or delivered to the customer) over time.
Dedicated Customer/Product Owner	Designating organizational member to prioritize the interests of the customer in long-term development process.
Coordination Mode	
Scrum Standup Meeting*	Discussing orally the organizational priorities and tasks in a short meeting, often while standing up.
Kanban Board, Story Mapping	Depicting visually the goals and tasks at various stages of a product or service development process, often on a whiteboard with sticky notes.

Practices indicated by * are included in our study of iterative coordination.

Figure 4.1

Figure 4.1 *(previous page)*: **Managerial Practices Associated with Agile.** Summary of most popular specific practices associated with or derived from the Agile management framework. Organizations considered to be Agile use at least one but not necessarily all of these specific practices. Only most popular practices in 2020 are included here (VersionOne 2020). Practices indicated by * are included in our study of iterative coordination. The other practices not marked are not a formal part of this research but are included to illustrate distinction between iterative coordination and other ways that practitioners might implement Agile. We provide an organizing framework that categorizes these practices by their specific purpose: setting *Interim Deadline Frequency*, specifying goal *Prioritization Discussion Content*; and facilitating communication with a *Coordination Mode*. The types of *Prioritization Discussion Content* facilitate discussion on a pre-existing goal, an interim goal, and a long-term goal. These three categories roughly map to the three discussion questions, respectively, that we vary in our experimental intervention.

Iterative coordination is implemented in the form of frequent meetings to coordinate individuals on their multiple goals, held once a day or even multiple times a day. In each of these meetings, iterative coordination allows an organization to prioritize among their multiple goals via discussion questions, generally three. In a common formulation—which we apply in our experimental studies—each meeting features discussion centered around the following three questions: (1) "What have you accomplished since the last meeting?"; (2) "What are your goals until the next meeting?"; and (3) "What are your goals for the end of the project (and have they changed)?"[37] Question 1 updates members of an organization about prior work towards pre-existing organizational goals. Question 2 defines goals for the organization until its next meeting. Question 3 prompts members of an organization to revisit its overall set of goals. Question 3 especially differentiates iterative coordination from a regular meeting by forcing the organization to revisit its priorities with respect to its shared goals. Through this third question, an organization can revisit and potentially prioritize among multiple performance goals.

In the absence of iterative coordination, how would an organization and its members manage its multiple goals, all other factors held equal (e.g., organizational structure, re-

[37] In Agile practice, projects are time-bound within "sprints," which break up software development cycles into smaller chunks of time. In Question 3, the use of the word "project" refers to the end of an Agile sprint.

sources, etc.)? An extensive body of prior work documents why members of an organization have limited impetus to discuss their shared organizational goals. Should members of an organization choose to meet, they may update each other on prior tasks and decide which tasks to complete before the next meeting—but it is less likely that a simple meeting in and of itself would prompt iterating on its higher-level organizational goals (Lifshitz-Assaf et al. 2020). First, the mentally taxing, deliberate cognition needed for long-term goal-setting does not represent the natural disposition of most workers (Pervin 1992, Seijts et al. 2004, Critcher and Ferguson 2016, Ghosh 2020). Second, members in an organization naturally seek to avoid potential conflict when discussing their priorities among multiple organizational goals (Cyert and March 1963, Gavetti et al. 2012). Without a forcing mechanism for discussion as presented by iterative coordination, there is less impetus to revisit and prioritize among its higher-level shared goals. In summary, in the absence of iterative coordination, we argue that an organization revisits its goals less frequently, limiting opportunities for reprioritization.

4.2.2 Applying Iterative Coordination to Manage Innovation

Increasingly, organizations use iterative coordination to manage the pursuit of innovation, ranging from the development of new software applications to creating new broadcast programming (Rigby et al. 2018). Despite its widespread use in the management of innovation, it remains unknown whether and how managing goals with iterative coordination affects the outcomes of novelty and value. We now turn to how the previously undertheorized phenomenon of iterative coordination may shape the prioritization of the distinct goals of novelty and value when innovating.

Although basic intuition would suggest that an organization could choose to pursue either novelty or value (or both) using iterative coordination, we argue that practicing iterative coordination drives processes that ultimately result in the *implicit* prioritization of value over novelty. The prioritization is implicit since it does not involve an ex ante choice to favor value over time. This occurs due to two mechanisms: additional interim deadlines and goal reprioritization.

The first mechanism—additional interim deadlines—promotes value at the cost of novelty. Under iterative coordination, organizations meet more frequently to discuss shared goals. This creates additional interim deadlines for work (Gersick 1988, 1989, Waller et al. 2002). Prior literature suggests that deadlines serve as an impetus to integrate an organization's existing knowledge (Okhuysen and Eisenhardt 2002). For example, an approaching deadline might inspire a mobile application developer to combine the payment system she has been working on with the user interface for the rest of the application. In doing so, the organization realizes value from integration: the payment system would not contribute any value if not integrated with the rest of the application. This example highlights the primary purpose of integration in organizations, which is to realize value from the existing knowledge of its members (Grant 1996). For instance, the organization can integrate different individual perspectives to refine ideas and proposals that promise to deliver the most value (Girotra et al. 2010, Keum and See 2017).

Absent iterative coordination to prompt the revisiting of shared goals, organizations integrate less knowledge from their members. As individuals create new knowledge, they must share it in a way that is accessible to other members in an organization (Nonaka 1994, Spender and Grant 1996). Accordingly, when lacking the impetus to share knowledge through an intervention such as iterative coordination, organizations have fewer opportunities to integrate their knowledge in a way that creates value for the organization (Okhuysen and Bechky 2009).

Meanwhile, iterative coordination's implicit focus on integration drives out individual specialization needed to generate new knowledge, which may limit the emergence of novelty (Cyert and March 1963). Individual specialization develops and expands the bounds of an organization's knowledge, helping organizations identify novelty (March and Simon 1958, March 1991, Kaplan and Vakili 2015). When individuals specialize in their knowledge creation efforts, the organization can more efficiently identify and develop ideas that represent truly novel breakthroughs (Csikszentmihalyi 1996, Taylor and Greve 2006). But as individ-

uals focus on integrating their existing knowledge under iterative coordination, they leave less time for specialization. Given limited resources and attention, individuals may integrate or specialize, but they cannot pursue both processes simultaneously (Knudsen and Srikanth 2014). For example, the time it takes to communicate findings and integrate knowledge directly takes away from the time an individual has to specialize in their own work and develop new knowledge (Knudsen and Srikanth 2014, Levinthal and Workiewicz 2018).[38] In light of this well-documented trade-off between integration and specialization in organizations (Lawrence and Lorsch 1967), we would expect specialization and, ultimately, novelty to suffer as a result of iterative coordination.

While the pressure of additional deadlines may favor value as opposed to novelty in output, how can organizations avoid failing to meet originally stated goals of both high novelty and high value? To this end, a second mechanism of goal reprioritization becomes relevant. Hu and Bettis (2018, pg. 886) propose that an iterative approach to goal-setting "may sometimes, perhaps often, involve lowering expectations for one or more goal levels, especially under time pressure." Here, iterative coordination allows an organization to reprioritize among simultaneous goals of high novelty and value. With additional deadlines driving an organization toward the goal of value, we predict that a treatment of iterative coordination will lead an organization to reprioritize the balance between value and novelty in its goals, favoring the ultimate realization of value over novelty in its final outcome. We predict this to be true regardless of an organization's starting point in terms of its balance between value and novelty, as empirical work demonstrates that it is much more difficult to re-introduce novelty than value during the innovation process (Berg 2014).

Using this theoretical viewpoint as a starting point for our exploratory inquiry, we now describe the empirical findings from a series of experimental studies that cumulatively guide

[38] This trade-off is especially salient in our empirical setting of software development, which involves multiple specialist coders who attempt to create a novel and valuable software application. They must divide and allocate their time and attention to either integrating with each other's existing knowledge or to specializing on their own to create new knowledge.

the development of additional theory on iterative coordination's effects on innovation.[39]

4.3 Primary Study: Software Development Field Experiment

To study the effect of iterative coordination in innovation, we design and deploy a field experiment. Given iterative coordination's roots in the software industry (Sutherland and Sutherland 2014, Rigby et al. 2016b), we focus on the context of managing software development. To maintain managerial relevance, we sought an externally valid experimental context to demonstrate the impact of iterative coordination as a managerially implementable practice. We begin by presenting background on the externally valid empirical context—a software development competition, known generally as a hackathon—followed by the exposition of our procedure to administer iterative coordination as experimental treatment.

4.3.1 Experimental Setting

We partnered with Google LLC (Google), a multinational information technology firm, to embed a field experiment within a one-day software application ("app") development competition, or hackathon, hosted on the campus of a university in the northeastern United States.[40]

Hackathons: Background and External Validity

Over the last decade, hackathons emerged to play a pivotal role in software development culture and practice (Broussard 2015, Leckart 2015, Pan Fang et al. 2020). Hackathons commonly entail sets of software developers who compete in a contest to develop and present working software by the end of a timeframe of a day or two (Leckart 2012). Although some hackathons focus on particular themes or interest areas, they generally operate as open-ended design contests which embrace ambitious innovation goals. In spite of the short time frame allotted, what motivates participants at a hackathon are clear articulations of their project goals (Lifshitz-Assaf et al. 2020).

[39] As we are interested in developing new theory based on the phenomenon of iterative coordination, we avoid a formal statement of hypotheses, which are more appropriate for empirical studies of mature bodies of theory (Edmondson and McManus 2007). Instead, we offer theoretical predictions to guide the interpretation of results from our primary study in Section 4.3.

[40] Google LLC is the largest subsidiary of Alphabet Inc.

The competition aspect of the hackathon typifies the dynamic and entrepreneurial settings in the software industry where firms implement iterative coordination in practice (Ott et al. 2017). Each set of participants mirrors the composition of an archetypal software start up firm in terms of skills and size. In fact, many successful startup firms began as hackathon projects: the popular messaging app GroupMe was conceived at the 2010 TechCrunch Disrupt hackathon and acquired a year later by Skype for about $80 million (Arrington 2010, Ante 2011). Given these contextual factors and theory on what defines a firm, we refer to competing teams at a hackathon as firms.[41]

The participating firms compete against each other in a "market," where customer choice is represented by the evaluation of event judges. These judges evaluate the output of each firm at the end of the competition, rewarding selected firms with prizes based on a number of pre-selected criteria. Hackathons across contexts favor novelty and value in ideas and solution approaches, even as specific judging criteria may differ. This hackathon environment is well-suited to study iterative coordination: much like the features of software markets in which iterative coordination is adopted as a management practice, the hackathon environment prioritizes innovation.

Hackathon sponsors commonly provide mentors to participating firms throughout the competition. Given their non-evaluative, authority-free support role at hackathons, mentors are ideal facilitators of our iterative coordination treatment.

Competition Specifications

In terms set by Google, competing firms developed a software application that provided an innovative solution to some social objective, e.g., a sustainability app to track personal carbon footprint or an app for NGO fieldworkers to collect data. Each firm defined its organizational goal as a description of a novel and valuable technical product that they

[41] A firm, or more broadly an organization, is defined as a bounded system where more than one agent shares system-level goals and where each constituent agent is expected to make a contribution. In this conceptualization, "boundaries and goals jointly identify organizations uniquely" (Puranam et al. 2014, pg. 164). Using this definition, Puranam et al. (2015, pg. 381) note that "there is no basis (besides convention) on which one can say that a three-person firm is an organization but a four-person team is not; to the extent these are goal directed multi-agent systems, they are both organizations."

wished to create by the end of the competition.[42] Consistent with standard hackathon practice, firms chose the specific problem they wished to work on, provided it was in service of the general theme of the event. Prizes totaling $2,000 USD in monetary value were provided by Google to top-performing firms.[43]

In collaboration with Google, we recruited firms consisting of software engineers to compete in the hackathon.[44] Competing firms were composed of upper-level undergraduate computer science majors from local universities and professional and freelance software developers. Individual participants qualified based on their prior collaborative software development experience, assessed through a submitted portfolio of past projects. Participants registered together in firms of two to four members in a pre-event survey designed with our co-sponsor; the pre-event survey data also served as a source of control variables and for screening potential participants on the technical skills necessary to be productive during the hackathon.

Prior to the start of the competition, firms were randomly assigned into treatment and control conditions. In all, 38 firms competed in the hackathon, consisting of 112 participants (62 students and 50 professionals).[45]

Although firms had flexibility to define the nature of their applications, they were required to meet a few basic requirements for the competition. First, they were required to use a fixed development toolkit provided by Google. This finite software toolkit limits the product attributes firms can consider and use to build their applications (Levinthal 1997, Fleming 2001). By holding available technological inputs constant across treatment and control conditions, we strengthen our ability to interpret the causal effect of our intervention. Second,

[42] Scholars have long recognized the potential limitations to the concept of an "organizational goal" (Cyert and March 1963, Simon 1964, Gaba and Greve 2019). For instance, while phrased as a singular objective, an "organizational goal" often encompasses multiple demands that must be simultaneously satisfied (Simon 1947, Gaba and Greve 2019). This is especially salient in our context of the development of new technology, where managing multiple performance objectives simultaneously is a defining characteristic (Hu and Bettis 2018).

[43] Academic research grants supported the operational expenses of the experiment, event, and venue.

[44] Online Appendix C.1 documents the materials used to recruit participants for the hackathon.

[45] Online Appendix C.1 presents a post hoc analysis of statistical power.

to collect the detailed data over time on development processes, all firms were required to record their work over the course of the competition with the open-source version-control software, Git. By tracking the emergence of software code, Git allows for the detailed measurement of software development activities over time.[46] Finally, Google communicated the need for solutions that would be both novel and valuable to customers—thereby articulating the multiple goals of innovation.[47] Each of these requirements was clearly communicated to all hackathon participants in an opening presentation prior to the official start of the competition.

Within the scope of these requirements, at the start of the competition, firms set a fairly diverse set of goals for themselves to pursue. For example, one firm wanted to build a virtual reality application that would use facial recognition technology to help senior citizens with dementia or Alzheimer's disease identify friends and family. Another firm wanted to build a mobile application with a proprietary algorithm to match refugees with support communities and resources. Yet another firm wanted to create an application that would use artificial intelligence to automatically categorize unstructured data inputs from fieldworkers of small NGOs to make the data practically usable later.

4.3.2 Experimental Procedure

Experimental Treatment

Leveraging the natural features of the hackathon format, Google engineers who served as mentors to each firm facilitated iterative coordination. At the start of the treatment period, all firms, in both the treatment and control conditions, were approached every two hours by their randomly assigned mentor, who was instructed to offer a null greeting in reference to an item on the schedule (e.g., "How was lunch?"). Each mentor appeared before an equal number of treatment and control firms. After this greeting, mentors visiting control firms concluded their interaction. In contrast, mentors visiting treatment firms would facili-

[46] The Git, specifically GitHub, interface allows us to see which member contributed to which portion of the project over time. As each member writes code, they submit it to the shared GitHub repository that represents the body of code for the overall project by the firm.

[47] Online Appendix C.1 presents an example of Google materials communicating this guidance.

tate a short iterative coordination meeting asking treatment firms to discuss three questions: (1) "What have you accomplished since your last check-in?"; (2) "What are your goals until the next check-in two hours from now?"; and (3) "What are your goals for the end of the day (and have they changed)?"[48] Per instruction, mentors did not provide any quality judgments to firms during these iterative coordination meetings; rather, they simply served as facilitators for group discussion.[49] To ensure that the treatment closely reflected practice, we devised the three aforementioned questions after observing iterative coordination meetings used at Google; we further verified the external validity of these questions in interviews with other engineers from Google and peer firms that practice iterative coordination.[50]

We built in a pre-treatment period of 2.5 hours in which no firms were treated. The inclusion of this pre-treatment period allows us to include firm fixed effects and run a generalized difference-in-differences regression model. As we shall address in our firm-minute analysis of the firm processes in Section 4.3.4, the firm fixed effects control for time-invariant quality differences between firms, bolstering causal identification in case there was any further unobserved time-invariant heterogeneity between firms not addressed by randomization of the treatment. After this pre-treatment period, the periodic iterative coordination meetings occurred every two hours until the close of the competition for treatment firms but not for the control firms. Treatment firms experienced three iterative coordination meetings over the course of the competition.

Other Experimental Design Considerations

To ensure causal identification from the field experiment, we made several explicit efforts to limit the participant perception of mentor authority, to prevent participant awareness of heterogeneous treatment, and to ensure that the mentors properly administered the

[48] In the final check-in, two hours before the end of the competition, only the first and third questions were asked.

[49] We instructed mentors to provide only *technical* feedback if *directly requested*. Mentors in both the treatment and control conditions restricted themselves in this way and did not proactively speak to firms outside of the formal interventions. By limiting feedback in this way, we exclude any normative guidance on the value or novelty of the product being developed.

[50] Peer firms included Twitter, Inc. and Cisco Systems, Inc.

iterative coordination treatment.

We minimized the perception of mentors as authority figures in three ways. First, it was clearly communicated to the participants that mentors absolutely did not serve as or communicate with the judges in the competition. Second, the mentors were demographically similar (e.g., age, professional background, etc.) to the participants, minimizing perceptions of authority enforced by differences in social status (Ashforth and Mael 1989, Lincoln and Miller 1979). Third, the mentors did not provide any unsolicited normative guidance to participants.

Participants remained unaware that treatment firms and control firms experienced different mentor interactions through a number of design decisions made in conjunction with our partner Google. The undesirable consequences of such awareness range from spillover effects from treatment to control (Duflo and Saez 2003), to Hawthorne effects where firms act differently due to their awareness of being observed for study (Levitt and List 2011). First, we physically separate the workspaces of treatment and control to minimize the chance of across-condition interaction that might lead to awareness of different attention from mentors. Second, to reinforce perception of parity, the assigned mentors visit firms every two hours in both treatment and control, as previously noted. While a pure counterfactual control to our iterative coordination treatment could conceivably involve no interaction with mentors every two hours in control, excluding control firms from any mentor visits would risk increasing awareness of differential mentor attention. In addition, keeping the mentor visit events uniform for both treatment and control firms has the desired effect of keeping constant the time for cycles of software development work. Discussions with Google made it evident that any mentor interaction could break up software development cycles in a way that may otherwise not occur. To isolate the causal effect of iterative coordination questions—without the potential confounding factor of different cyclicality—we ensured that both treatment and control would be visited by mentors on the same two-hour cycle. Online Appendix C.1 further details the experimental design—including detailed floor plans of the physical space

and mentor scripts—as well as several precautions taken to ensure that Google mentors properly administered the treatment of iterative coordination.

We now discuss the data, statistical methods, and results of the field experiment looking at the effect of iterative coordination: first, a section on innovation outcomes, i.e., value and novelty; second, a section on potential processes, i.e., integration and specialization, which may lead to innovation.[51]

4.3.3 Organizational Outcomes

We begin our analysis of iterative coordination on innovation by analyzing its effects on outcomes of *Value* and *Novelty*, two key dimensions of innovation (Amabile 1983, Kaplan and Vakili 2015).

Data

Our dataset to study performance outcomes consists of a cross-section of firm project evaluation by expert judges after the end of the hackathon competition, combined with a set of covariates to serve as control variables collected through a pre-event survey.

Each firm was visited by a third-party panel of three judges to evaluate their projects at the end of the competition. Given that novelty and value in innovation are socially constructed by perceptions (i.e., novelty is established relative to existing products), expert judge evaluations are the appropriate method for measuring innovation (Amabile and Pratt 2016). These judges were not involved with and were unfamiliar with our study design. Each judge had several years of work experience in the software industry and had both participation and judging experience in other hackathons prior to our event. The judges tested and interacted with the applications that the firms developed.

As part of the formal registration process for the hackathon, participants were asked

[51] As a necessary condition to conduct the primary field study, the researchers and their university executed a contract with the corporate partner on the study that specifies requirements for non-disclosure of individually identifiable data. These contractual terms intend to protect the privacy and intellectual property of the partner's employees (i.e., mentors), third-party volunteers (i.e., judges), and prospective employees and independent software vendors (i.e., participants). These terms restrict the public release or sharing of data from this study. The data from the follow-on laboratory study does not have this type of contractual restriction and is made publicly available.

Table 4.1: **Primary Field Study: Variable Definitions and Summary Statistics of Firm Outcomes from Judge Evaluation.** Judges were asked to score each firm's final submission on a 1 to 5 Likert scale according to the criteria provided by our corporate co-sponsor Google.

Variable	Definition	Mean	SD
Value	How much does your project appeal to the intended market? (Likert scale 1 to 5)	2.553	1.796
Novelty	Does the project help solve the problem in a new and ambitious way? (Likert scale 1 to 5)	2.316	1.726

to complete a short registration survey that was designed with our co-sponsor. We use this survey data to generate firm control variables and assess the efficacy of experimental randomization.

Variables

To measure outcomes of innovation, we use two measures capturing different dimensions of innovation for the applications developed by firms. Our first outcome measure is *Value*, which measures the extent to which an application caters to its existing target customer base. Our second dependent variable, *Novelty*, captures whether an application solves customer problems with a new approach within the scope of Google's furnished app development toolkit. Judges scored each firm's final project along the two aforementioned outcome categories based on a Likert scale of 1 to 5, summarized in Table 4.1. The use of these specific criteria to evaluate software applications had been validated by our co-sponsor, Google, from experience hosting prior hackathons. Furthermore, value and novelty capture independent components of innovation (Amabile 1983, Singh and Fleming 2010, Kaplan and Vakili 2015), as discussed earlier. We also conduct an independent validation of these measures, detailed in Online Appendix C.2.

To control for observable time-invariant across-firm heterogeneity that might remain despite the randomization process, we include several firm characteristics drawn from the pre-event survey as independent variables. *Current Student* is the firm mean of student status (with students taking a value of 1; and 0 otherwise). *Graduate Degree* is the firm mean

Table 4.2: **Primary Field Study: Firm Characteristics and Correlations.** Means and standard deviations in parentheses for firm-level observations of the full sample ($N = 38$), treatment condition, and control condition. The Difference column shows a t-test of difference in means between the treatment and control condition, with standard errors in parentheses. The numbered columns to the right display pairwise correlations.

	Sample				Pairwise Correlation						
Variable	Full	Treatment	Control	Difference	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Current Student	0.539	0.495	0.579	0.0838	1						
	(0.386)	(0.402)	(0.377)	(0.126)							
(2) Graduate Degree	0.360	0.375	0.346	-0.0292	-0.253	1					
	(0.344)	(0.334)	(0.361)	(0.113)							
(3) GitHub	0.901	0.954	0.854	-0.0995	0.227	-0.557	1				
	(0.192)	(0.138)	(0.223)	(0.0609)							
(4) Google Development	0.461	0.463	0.458	-0.00463	0.496	0.155	-0.049	1			
	(0.323)	(0.363)	(0.292)	(0.106)							
(5) Software Development	3.695	3.838	3.567	-0.271	-0.411	0.359	-0.124	-0.006	1		
	(3.839)	(3.661)	(4.083)	(1.264)							
(6) Prior Hackathons	1.825	1.662	1.971	0.309	0.104	-0.079	-0.033	0.238	0.349	1	
	(1.186)	(0.943)	(1.378)	(0.387)							
(7) Firm Size	2.947	2.722	3.150	0.428	0.132	-0.12	0.009	0.242	-0.361	-0.214	1
	(0.837)	(0.826)	(0.813)	(0.266)							

of educational experience (with Master's and Doctoral degrees taking a value of 1; and 0 otherwise). *GitHub* is the firm mean of prior history using GitHub. *Google Development* is the firm mean years of experience with the development toolkits provided by Google. *Software Development* is the firm mean years of professional software development experience. *Prior Hackathons* is the firm mean of hackathons attended prior to the event. *Firm Size* is a count variable of the number of members in the firm.

Table 4.2 presents the summary statistics and correlations of these control variables. In addition, we use these same variables in t-tests of differences across treatment and control firms as a check of the efficacy of the randomization, also shown in Table 4.2. We do not find any evidence of systematic bias in our randomization. [52]

[52] As an additional check on the efficacy of randomization, we obtain and code data on the starting goals of the firms participating in the hackathon. By starting goal, we mean the initial description of a novel and valuable technical product that they wished to create by the end of the competition. We find no statistically significant difference between firms in the treatment condition and control condition in their *Starting Goal Value* or *Starting Goal Novelty*. The comparability of these measures across the two conditions provides further confidence in the efficacy of the experimental randomization. Online Appendix details these analyses C.2. We thank an anonymous reviewer for this suggestion.

Estimation Model

To compare end-of-competition firm outcomes, we run cross-sectional OLS models with dependent variables for the evaluation categories regressed on an indicator variable for treatment, with firm control variables drawn from the pre-event survey listed in Table 4.2. In addition, we include the dummy indicator *No Evaluation* to control for whether a firm officially submitted an application for judge evaluation, which commenced a half-hour after the competition officially closed. Regardless of participation in judge evaluation, all firms nonetheless stayed to the end of the competition, and their project code was observable to us throughout the competition.

Results

Table 4.3 presents the effects of iterative coordination on final outcomes. Model 4.3-1 demonstrates that iteratively coordinating firms scored on average 0.614 points higher (0.341 standard deviations higher) on *Value* than informally coordinating firms ($p < 0.01$). Removing firms that did not participate in judge evaluation from our sample, Model 4.3-3 shows that iteratively coordinating firms scored an average of 0.846 points higher (0.471 standard deviations higher) on *Value* ($p < 0.01$). Supporting our findings of a positive association between iterative coordination and *Value* are Models 4.3-2 and 4.3-4, which include the full set of firm controls to control for observable heterogeneity not addressed by the experimental randomization.

In contrast, Model 4.3-5 indicates that iteratively coordinating firms scored approximately a half-point less than control firms (0.289 standard deviations lower) on *Novelty* ($p < 0.10$), with a similar negative association in Model 4.3-7 ($p < 0.10$). Models 4.3-6 ($p < 0.05$) and 4.3-8 ($p < 0.05$) demonstrate the robustness of this result when including the full set of firm controls.

Supplemental Analyses

We conduct a number of additional tests to verify the robustness of these findings to alternative specifications and explanations. We confirm the direction and statistical signif-

Table 4.3: **Primary Field Study: Regression Analysis of Firm Outcomes from Judge Evaluation.** Ordinary least squares (OLS) estimation of cross-sectional data at the firm level. Robust standard errors shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. In estimates involving the full sample, *No Evaluation* takes a value of 1 for firms that decided not to undergo the judging process.

	Value				*Novelty*			
	(4.3-1)	(4.3-2)	(4.3-3)	(4.3-4)	(4.3-5)	(4.3-6)	(4.3-7)	(4.3-8)
Treatment Condition	0.614**	0.546*	0.846**	0.661*	-0.499†	-0.692*	-0.687†	-0.960*
	(0.222)	(0.208)	(0.290)	(0.302)	(0.278)	(0.332)	(0.375)	(0.370)
Current Student		0.725*		0.991		0.145		0.638
		(0.353)		(0.650)		(0.390)		(0.793)
Graduate Degree		0.430		0.365		-0.458		-0.877
		(0.384)		(0.590)		(0.520)		(0.697)
GitHub		0.120		0.106		0.893		0.907
		(0.798)		(0.965)		(0.996)		(1.106)
Google Development		-0.207		-0.310		0.898†		1.006
		(0.410)		(0.813)		(0.498)		(0.695)
Software Development		-0.050†		-0.039		0.007		-0.003
		(0.028)		(0.041)		(0.040)		(0.048)
Prior Hackathons		-0.144		-0.182		-0.096		-0.046
		(0.087)		(0.107)		(0.124)		(0.162)
Firm Size		-0.113		-0.096		-0.222		-0.421†
		(0.102)		(0.149)		(0.202)		(0.223)
No Evaluation	-3.497***	-3.702***			-3.336***	-3.331***		
	(0.189)	(0.195)			(0.202)	(0.256)		
Constant	3.274***	3.590***	3.154***	3.456***	3.518***	3.284***	3.615***	3.782**
	(0.219)	(0.883)	(0.249)	(1.170)	(0.214)	(1.151)	(0.241)	(1.434)
R^2	0.874	0.922	0.261	0.574	0.774	0.824	0.117	0.454
Sample	Full Sample		Evaluation Only		Full Sample		Evaluation Only	
Observations	38	38	27	27	38	38	27	27

icance of the main effects across these analyses to rule out alternative explanations. First, we conduct a post hoc analysis of the effect of heterogeneity in the value and novelty of the articulated starting goal of each firm at the start of the hackathon. This post hoc analysis provides suggestive evidence that there is a boundary condition to the effect of the iterative coordination experimental intervention. Firms that start with a goal that is already high in value and/or low in novelty are still impacted by the intervention but to a lesser degree. We also use this data to show that the mean value and novelty of starting goals are comparable across the treatment and control conditions, supporting the efficacy of the experimental randomization. Relatedly, this post hoc analysis provides suggestive evidence that goals do shift due to iterative coordination based on the difference between a firm's starting goal and what it ultimately delivers at the end. Second, we find the same patterns in an ordered logit analysis mirroring Table 4.3. Third, we rule out the alternative story that there may be differences in productivity across firms by looking for possible differences in project completion by the end of the competition. Fourth, we do not find that the intervention has any effect on selection by firms into evaluation. Fifth, we confirm that the same pattern holds when we allow for firm characteristics as moderators. The Online Appendix presents the full results of these robustness checks in detail.

In measuring the effect of iterative coordination on innovation outcomes, we find the results are mixed: while iteratively coordinating firms develop products that are more valuable, these products are simultaneously less novel. Despite the competitive context of the Google hackathon demanding novel solutions, our results demonstrate that firms treated by iterative coordination questions produce less-novel output.

4.3.4 Organizational Process

Given the results outlined in the previous section, we ask: What processes might iterative coordination be influencing that associate with more valuable yet less novel output? In Section 4.2, we present a theory for iterative coordination that associates greater value and reduced novelty with the processes of integration and specialization, respectively. We

Table 4.4: **Primary Field Study: Variable Definitions and Summary Statistics of Firm Process from Software Code.** Dependent variables for firm process defined below with their conceptual interpretation. Observations are at the firm-minute level, with 20,520 firm-minute observations across 38 firms.

Variable (Interp.)	Definition	Mean	SD	Min	Max
Code Integration Action (Integration)	Count of actions taken by the firm to integrate software code into and within the firm's shared code base.	1.954	3.316	0	20
Advanced API Specialization (Specialization)	Count of uses of non-required specialized and advanced application programming interface (API) procedures, protocols, and tools.	0.700	1.474	0	7

now dive into firm software code to unpack iterative coordination's influence on processes of integration and specialization, both of which have long been studied by organizations scholars for their relationship to innovation (Lawrence and Lorsch 1967, Grant 1996, Levinthal and Workiewicz 2018).

Data: Software Development

To study the effect of iterative coordination on innovation process, we analyze a balanced firm-minute panel with dependent variables measuring integration and specialization in the software development process based on our minute-by-minute tracking of the firms' updates of their software code through Git.[53] With each update timestamped to the minute, our novel empirical strategy achieves a precise level of granularity.

Our dependent variables measure specific actions in the software development process consistent with integration or specialization, as summarized in Table 4.4.[54]

Dependent Variables

To measure organizational integration, *Code Integration Action* consists of the stock count of actions taken by the firm to integrate software code within the firm's shared code

[53] Git is a free, open-source version-control system that enables distributed software development. Version control keeps track of all the changes developers individually make to the firm's "repository" of source code for a project. Git archives each firm's source code repository at each update. Should errors be made during the development process, a developer can easily restore the firm's repository to that of a prior "commit" or update, which stores a snapshot of the firm's repository at the time the update was made.

[54] Online Appendix C.3 presents the matrix of pairwise correlations.

base. This measure captures two types of convergent development efforts in facilitating an integrated codebase. First, individual software developers, who may specialize and write some code independently, must combine their individual code with the firm's shared code base to integrate it with the overall project. Second, developers may combine code that is already in the shared codebase, thereby further integrating aspects of the project. The required version-control software enables and tracks these integrative activities, which are a standard part of software workflow management.

To measure specialization, *Advanced API Specialization* measures a firm's use of non-required specialized and advanced application programing interface (API) procedures, protocols, and tools in their codebase. Although firms were required to use a broader toolkit provided by our sponsor Google, there were several advanced API tools available to the developers for free that were encouraged but not required in the competition. These tools allow firms to use a number of advanced cloud-based features: analyze data using artificial intelligence capabilities, conduct natural language processing, leverage remote graphics processing units (GPU) for machine learning and 3D visualization, and connect with internet-of-things (IOT) devices, among other functionality. We measure the use of these tools by identifying the number of API calls or requests to these tools appearing in a firm's codebase. These tools require in-depth specialized knowledge to use, beyond the common knowledge developers would have coming into the competition. Moreover, these tools were only free in the context of our competition; they were available as paid enterprise software outside of the competition, making it likely that developers would not use them regularly prior to the competition. An optional tutorial on these advanced features was available to all hackathon participants.[55] Because *Advanced API Specialization* reflects advanced technical development beyond the expected standards, we use it to measure specialization.

[55] An additional requirement of Google LLC's co-sponsorship of our event was the inclusion of a tutorial on advanced features of one of the competition's required app development toolkits. This was offered to all firms late in the day, immediately after the second mentor check-in, and attendance was optional.

Estimation Model

We use these two dependent variables in a firm-minute panel to estimate the following differences-in-differences model:

$$Y_{it} = \beta(Treatment_i \times Post_t) + \alpha_i + \delta_t + \epsilon_{it}.$$

Y_{it} represents the dependent variables of *Code Integration Action* and *Advanced API Specialization*. $Treatment_i$ is an indicator variable taking a value of 1 for firms treated by iterative coordination, and $Post_t$ is an indicator variable equaling 1 after the completion of the first of three mentor check-ins. Our coefficient of interest β estimates the effect of iterative coordination meetings on Y_{it}.[56] The intentional inclusion of a pre-treatment period in the experiment allows us to include firm fixed effect α_i to control for time-invariant unobserved confounding factors (e.g., the complexity of the firm's chosen problem, Google toolkit know-how, etc.).[57] δ_t is a minute fixed effect to control for potential shocks across all firms during the hackathon (e.g., the beginning of lunch service at the event). We cluster robust standard errors at the firm level.

Results

Table 4.5 reports the results of regression analyses that test the effects of iterative coordination on integration and specialization. Model 4.5-1 reveals that treatment is positively and statistically significantly associated with *Code Integration Action*. That is, after iterative coordination meetings commence, iteratively coordinating firms conduct on average 2.074 more code integrations than control firms.

On the other hand, Model 4.5-2 displays a negative and statistically significant relationship between iterative coordination and *Advanced API Specialization* use. Specifically,

[56] $Treatment_i$ and $Post_t$ were not independently estimated in the model because they are collinear with the more precise fixed effects of α_i and δ_t, respectively.

[57] Following the best practice of recent field experimental research (e.g., Chatterji et al. 2019), we took a conservative approach using firm fixed effects to control for the chance possibility that there might be lingering unobservable time-invariant firm-level heterogeneity even after randomization. These firm fixed effects subsume all of the control variables that we use in the prior analysis of firm outcomes, e.g., *Graduate Degree*. We need a pre-treatment period to include firm fixed effects to ensure econometric identification of the key coefficient, which is the effect of treatment in the post period, or the coefficient on $Treatment_i \times Post_t$. When we include fixed effects, $Treatment_i$ is collinear with firm fixed effects and thus omitted from the regression estimation because it cannot be identified.

Table 4.5: **Primary Field Study: Regression Analysis of Firm Process from Software Code.** Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	(4.5-1) *Code Integration Action*	(4.5-2) *Advanced API Specialization*
Treatment x Post	2.074*	-1.124**
	(0.878)	(0.408)
Firm FE	Yes	Yes
Time FE	Yes	Yes
R^2	0.456	0.335
Adjusted R^2	0.441	0.317
Firms	38	38
Observations	20,520	20,520
Level	Firm-Minute	Firm-Minute

iteratively coordinating firms conduct 1.124 fewer highly specialized uses of Google's advanced application development toolkit in the post-period.

We demonstrate these results graphically in Figure 4.2, which plots the estimates of the difference in integration and specialization actions in the hours before and after the start of treatment. Across both graphs, we find no significant differences across treatment and control firms. This offers additional support for the efficacy of the randomization of treatment.

Supplemental Analyses

We devise a number of additional tests to assess the robustness of these findings. First, we devise two additional measures of integration and specialization based on the underlying structure of the file hierarchies in the software code. These two measures are based on branching factors, a standard performance measure in the computer science literature (Knuth and Moore 1975, Baudet 1978, Muja and Lowe 2009). We find statistically significant results consistent with those reported in Table 4.5. We carry these measures through the other robustness checks.

Second, to ensure the robustness of our results relative to estimates of standard errors

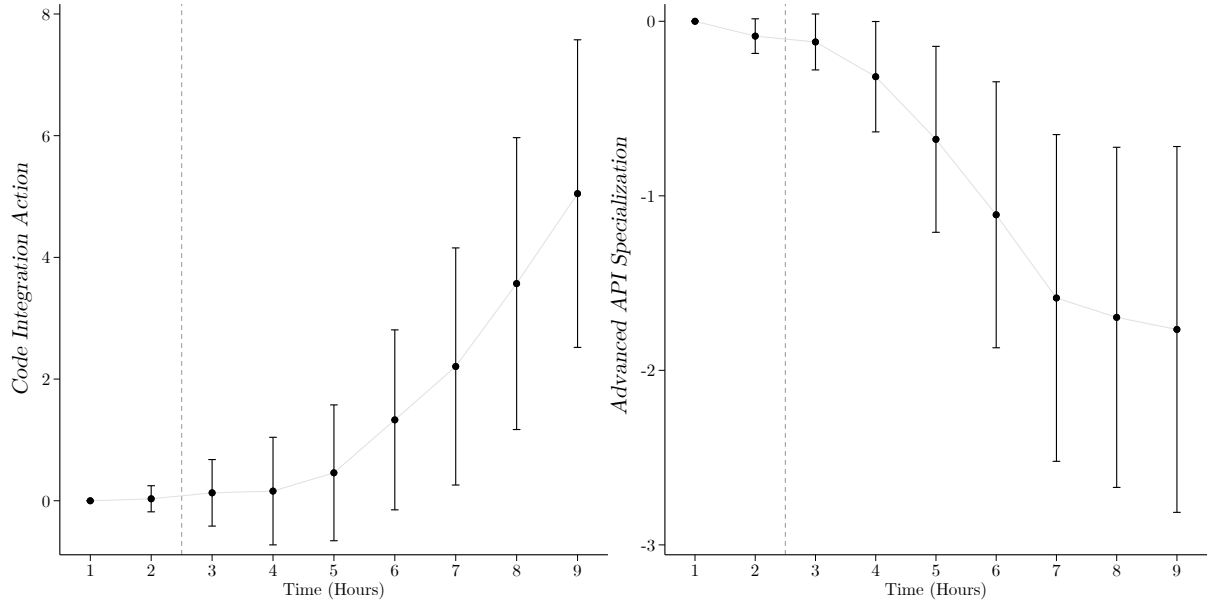


Figure 4.2: **Primary Field Study: Effect on Process over Time.** These four graphs depict the effect of the iterative coordination treatment over time. Each point estimate represents the difference in the level of the dependent variable between the treatment and control groups. The dotted grey line indicates the start of the treatment period. These graphs were constructed based on OLS regression estimates using an indicator variable for each hour of the experiment interacted with the treatment variable. The indicator variable for the first hour was necessarily omitted for estimation tractability, but shown here as the baseline term, set at a value of 0, i.e., equating the treatment and control groups. Firm and time (minute) fixed effects included. The 95% confidence intervals shown derive from robust estimates of standard errors clustered at the firm level.

which may be underestimated due to serial correlation in long time-series panels, we follow Bertrand et al. (2004) and run our main analysis with observations collapsed to a pre- and post-period. We find statistically significant effects consistent with our findings in Table 4.5.

Third, given the cumulative nature of our treatment—three administered iterative coordination meetings—we assess iterative coordination's cumulative influence on integration and specialization after each iterative coordination meeting. A concern for the viability and interpretability of our results would arise if, for instance, iterative coordination's effects on integration and specialization were observed early in the post-period and were not sustained through the rest of the total observation period. As we would expect, our estimates of the effect of iterative coordination over time identify larger (further from zero) point estimates with greater statistical significance in later periods.

Fourth, we consider the extent to which observed differences between iteratively coordinating firms and firms in control may be driven due to differences in underlying productivity. If, for instance, iterative coordination hindered productivity, this alternative mechanism may instead explain iterative coordination's negative relationship with specialization. Nonetheless, we find that iterative coordination bears no significant effect, positive or negative, on a firm's raw productivity, mitigating this alternative explanation.

Fifth, we find no statistically significant difference in the amount of time that treatment and control firms spent in their mentor interactions or afterwards to regroup and get back to work. The time spent for these interventions is very short compared to the overall length of the competition.

Sixth, we assess the results relative to potential moderating firm characteristics. The full results of these six categories of additional analyses are provided in the Online Appendix.

4.3.5 Discussion of Primary Study

We now holistically consider how the empirical findings of the primary field experiment fit together. We first show that iterative coordination affects the innovative output of the firms we study, where it associates positively with value and negatively with novelty as judged

in the final products of each firm. We then turn to the granular software code data to understand what processes iterative coordination might impact. Consistent with our theory, we find that iteratively coordinating firms take more development actions oriented towards integration, as measured by changes in *Code Integration Action*. Meanwhile, they invest less in advanced, novel uses of Google APIs, suggesting decreased specialization. To formally test an empirical connection between specialization and novelty, we perform a post hoc mediation analysis, reported in the Online Appendix.[58] This analysis suggests there is reason to believe that integration and specialization are potential mediators that link iterative coordination to value and novelty, respectively. Nonetheless, we recommend caution in broadly interpreting this particular post hoc finding: we do not exogenously vary mediators of *Code Integration Action* and *Advanced API Specialization*, which would represent the ideal empirical design for a causal mediation analysis. Overall, the processes of integration at the cost of specialization help illustrate goal reprioritization towards value over novelty—as firms shift their focus towards value over novelty objectives, the drive for individuals to specialize and create new knowledge to identify novel outcomes becomes less salient.

Although the results thus far explore iterative coordination's influence on innovation outcomes and related processes of integration and specialization, they leave open questions regarding mechanism. For instance, what provides the impetus for greater integration and value in iterative coordination? We previously theorize in Section 4.2 that (1) the addition of interim deadlines and (2) opportunities for reprioritizing among goals would provide an impetus for integration and the shift towards value over novelty in outcomes. Indeed, our post hoc analysis of starting goals suggests that iterative coordination generally pushes firms away from a starting goal of novelty and towards value as realized in their final output. Importantly, this empirical finding is consistent with our theoretical perspective that allowing for reprioritization of goals leads to value over novelty in output.

To more directly unpack the influence of the aforementioned mechanisms of interim

[58] Online Appendix C.2 presents the post hoc mediation analysis.

deadlines and opportunities for goal reprioritization, we conduct a follow-on laboratory experiment that allows for data to be collected to study these two mechanisms, beyond what was possible in our field setting.

4.4 Follow-On Study: Product Development Laboratory Experiment

To complement our primary field experiment, we run a second experiment in the laboratory to study the influence of iterative coordination's two mechanisms on innovation outcomes: interim deadlines and the opportunity to update and reprioritize goals. In addition to helping unpack these two mechanisms for iterative coordination, a follow-on lab experiment yields a number of desirable features. First, it follows the best practice of prior work to combine field data with laboratory experiments (Stoop et al. 2012). Although our primary field experiment provides the benefit of external validity and applicability, it represents a less-controlled experimental environment. The laboratory environment provides that control and precision. Second, we collect a broader set of data, not possible in the field, to build alternative measures that confirm our findings in the field and allow us to measure alternative mechanisms that may take place and to rule them out. In particular, we address the extent to which the interventions, in the form of formal meetings, account for time that participants would otherwise be working, and whether and how the intervention affects the degree of coordination that would otherwise take place in between meetings.

4.4.1 Experimental Design

In the experiment, teams design a new dormitory or apartment product concept for a manufacturer. The teams compete with other teams for a cash prize based on their proposed product. We implement this task based on prior work by Girotra et al. (2010).[59] We randomly assign teams to one of three experimental conditions described later, where we vary the implementation of the iterative coordination treatment. We pre-registered this labora-

[59] Online Appendix C.4 presents the full details of the task.

tory experiment online with the Open Science Framework.[60]

Participant Sample

We recruited 210 participants, drawn from the general population, to participate in this study in a behavioral research laboratory at a university in the northeastern United States. To arrive at this sample size, we use the effect sizes estimated in the primary field experiment in an a priori power analysis. Given the nature of our task, we imposed a requirement for our study that all participants must have an education level of at least some college education and a high-school diploma or equivalent. Online Appendix C.4 details the recruitment procedure and the a priori power analysis.

We randomly assign participants into 70 teams of three individuals and randomly assign each of these teams to one of three conditions.[61] Verifying the validity of this randomization, we find no statistically significant difference in individual characteristics across the three conditions.[62]

General Procedure

After entering the laboratory and completing a pre-experiment survey, participants receive instructions on their product development task. We randomly assign participants into teams of three and separately escort each team to their own private room to begin the experiment. Each team worked in their own private room for sixty minutes. We provide teams with sketch paper to make preliminary individual drawings; each individual could write on the sketch paper with a unique-color pen assigned to the individual. Each room contains a whiteboard for the team to illustrate its final product submission.

After the sixty-minute experiment, participants first complete a post-experiment sur-

[60] Pre-registration documentation available at: `https://osf.io/c7qmw/?view_only=e53498ecf5004ac695f830da6752497a`.

[61] Using the G*Power software (Faul et al. 2007), the a priori power suggested that a sample size of more than 42 divided across the three conditions would be appropriate. Because this varies depending on the effect of interest, we overshoot this number and use a sample of 70 in the actual follow-on laboratory experiment.

[62] To confirm the randomization, we confirm there are not statistically significant differences across the three experimental conditions in *Age*, *Gender*, *Graduate Education*, *Current Student*, *Any Experience*, and *Years of Experience*. Online Appendix C.4 presents further detail on this randomization check.

vey. Participants then vote on which team's final product concept, among those in their session, is their favorite; they are barred from voting for their own product. Within a session, we assign all teams to the same experimental condition, so no team is unfairly advantaged for the prize. In addition to $25 USD in compensation to participate in the study, the team that receives the most votes from peers in the session wins a prize of $10 USD per person.

Conditions

We design three experimental conditions, seeking to vary iterative coordination by the frequency of its meetings and the extent to which its discussion questions allow for a reprioritization of organizational goals. With these two dimensions of variation in mind, we sought an experimental design that would allow us to simultaneously explore both dimensions, while minimizing the number of experimental conditions to maximize the number of teams per condition for statistical power; team-level experiments are especially expensive given that they require, in our case, three times the number of recruited participants for the equivalent power of an individual-level study. Given this constraint, this follow-on study has no pure control condition as in the primary study.

To implement the experimental treatments, a member of the research team acts as a team mentor who visits each team intermittently to administer the iterative coordination meeting intervention(s).[63] The general structure and content of the mentor-participant interaction parallel what was used in the hackathon in the primary study, except for the questions discussed and the frequency of meetings.

In Condition 1, the baseline condition, we subject teams to only one intervention, asking only Question 1 of an iterative coordination treatment: after 20 minutes from the start of the experiment, the mentor asks one question to each team, "What have you accomplished since the last check-in?", which updates all the members of the team on progress towards a pre-existing shared goal. In Condition 2, teams only experience one intervention at 20

[63] To avoid deception, research team members introduce themselves as a member of the research team.

Table 4.6: **Follow-On Laboratory Study: Experimental Conditions.** This table summarizes the experimental intervention design in each of the three conditions. Different sets of questions at different times were posed by the mentor in each intervention.

Time Elapsed	Condition 1 (Baseline)	Condition 2 (Question Composition)	Condition 3 (Number of Interventions)
20 Minutes	What have you accomplished since the last check-in?	What have you accomplished since the last check-in? What are your goals for the end of the day?	What have you accomplished since the last check-in? What are your goals until the next check-in? What are your goals for the end of the day?
40 Minutes			What have you accomplished since your last check-in? What are your goals for the end of the day? (And have they changed?)

minutes into the experiment, as in Condition 1, but we vary the question composition to include the opportunity to reprioritize shared goals; the mentor also asks, "What have you accomplished since your last check-in?" and "What are your goals for the end of the day?" We do not ask them, "What are your goals until the next check-in?" since there is no next check-in and it would be redundant with the question on the goals for the end of the day. In Condition 3, teams experience two interventions, one at 20 minutes and one at 40 minutes, and they discuss all iterative coordination questions mirroring those posed in the hackathon field experiment. Table 4.6 summarizes the questions asked in each mentor/team interaction.

The comparison between Conditions 1 and 2 captures the mechanism of prioritizing between organizational goals; Question 1 only entails describing work achieved in pursuit of a pre-existing goal. In contrast, whereas Conditions 2 and 3 hold constant the mechanism of prioritizing between organizational goals, what differs is the frequency of meetings and thus interim deadlines: Condition 3 experiences an additional iterative coordination meeting at 40 minutes. In essence, one could view these conditions as representing low (1), medium (2), and high (3) levels of thoroughness or intensity in the implementation of iterative coordination.

4.4.2 Data and Measures

Table 4.7 details the source, construction, and interpretation of our empirical measures, organized by the construct they intend to measure.[64]

Table 4.7: **Follow-On Laboratory Study: Variable Definitions and Sources.** Measures drawn from each team's *Final Output* design, *Video Recording* of their working session, and their *Individual Sketches*.

Variable	Definition	Source
Outcomes		
Value	Usefulness of the final product (Likert 1–5). Average of two independent rater assessments.	Final Output
Novelty	Novelty of the final product (Likert 1–5). Average of two independent rater assessments.	Final Output
Process		
Time to Integrate	Time in seconds into the experiment until the team began working on the final product on the board based on draft individual sketches.	Video Recording
Individual Sketches	Count of pages of draft sketches generated by individuals. Reflects total productivity of individual specialized work.	Individual Sketches

We document the final product of each team, which was a product drawing on a separate whiteboard. These final products were rated by two independent raters on the dimensions of *Value* and *Novelty*, scored on the validated criteria and Likert scale (1–5) as in the primary study. Given high levels of inter-rater agreement, 0.86 and 0.80 respectively, an average of the two ratings for each dimension of product outcomes was used in analysis.

We collect and code the work output of each individual, over the course of the experiment, and each team, at the end of the competition. We separately track the preliminary design work done by each individual over the course of the experiment in the form of sketches on regular pieces of paper. We use these sketches to measure individual specialization, *Individual Sketches*, which is the count of the pages of draft sketches generated by individuals. *Individual Sketches* serves as a proxy for the total productivity of individuals on a team in their specialized work.

[64] The data and instruments for the follow-on laboratory experiment are available through the Open Science Framework portal: `https://osf.io/c7qmw/?view_only=e53498ecf5004ac695f830da6752497a`.

We record the video and audio of each team throughout the course of the competition. As a measure of integration, *Time to Integrate* assesses how quickly teams begin the process of integrating their individual work into the final product. *Time to Integrate* measures the time elapsed in seconds from the start of the experiment to when a team first writes on the whiteboard where they are required to report their final project submission. Writing on the whiteboard is a more integrative team activity, in contrast to individually drawn preliminary sketches on paper which reflect a more individually specialized activity. To build this measure, a research assistant watches the video and takes down the time stamp of the first moment when a dry-erase marker touched the whiteboard; there is no ambiguity in the coding process. Because teams use the whiteboard space for the final product, the action of writing on the whiteboard reflects integration of team knowledge into the final product. *Time to Integration* serves as a salient and behavioral (i.e., non-survey-based) indicator of integration that we can observe in the laboratory.[65] We also use the audio to generate text transcripts for several supplemental analyses.

4.4.3 Results

In Table 4.8, we report the summary statistics and cross-sectional analysis of the results of this laboratory study that compares differences in the means of the measures between Conditions 1 and 2 and between Conditions 2 and 3.[66] The direction and statistical significance of these findings are preserved when we instead use an OLS regression model that contains indicators for Conditions 2 and 3, where the relationships of interest would be the coefficient on the Condition 2 indicator and the t-test of differences in the coefficients on the indicators for Conditions 2 and 3.

[65] Teams were explicitly encouraged to start with preliminary individual work on the individual sketch paper and wait before committing to a final product on the whiteboard. Accordingly, the timing of using the whiteboard indicates how long it took for the team to finalize its decision about which product idea to focus on and use for the presentation slide (i.e., whiteboard drawing). Research assistants identified the moment when teams first use the whiteboard to confirm that the experimental design worked in the way we intended. In all cases, the first use of the whiteboard occurred only after individual sketches were completed and after the teams had one or multiple explicit conversations about which ideas on sketch paper should be combined and which idea to finalize and implement on the whiteboard.

[66] Online Appendix C.4 presents the correlation matrix.

Table 4.8: **Follow-On Laboratory Study: Summary Statistics and Cross-Sectional Analysis.** The first three columns contain the mean and in parentheses the standard deviation of teams in each condition. The last two columns compare Conditions 1 vs. 2 and Conditions 2 vs. 3, respectively, based on a t-test of the difference in means; the values reflect the difference in means and in parentheses the standard error, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Sample				Difference in Means	
	Full	Condition 1	Condition 2	Condition 3	1 vs. 2	2 vs. 3
Outcomes						
Value	3.386	2.971	3.444	3.739	0.473**	0.295*
	(0.602)	(0.594)	(0.535)	(0.414)	(0.165)	(0.140)
Novelty	3.190	3.551	3.208	2.812	-0.342†	-0.397*
	(0.692)	(0.729)	(0.612)	(0.540)	(0.196)	(0.169)
Process						
Time to Integrate	1579.6	1969.3	1572.8	1196.9	-396.4*	-375.9*
	(659.5)	(496.6)	(771.8)	(427.4)	(190.2)	(183.1)
Individual Sketches	4.171	5.304	4.208	3.000	-1.096*	-1.208**
	(1.579)	(1.329)	(1.560)	(0.853)	(0.424)	(0.369)

Outcomes

We demonstrate that both the ability to prioritize among innovation goals and the frequency of meetings have a statistically meaningful effect on the outcomes, consistent with two suggested mechanisms of iterative coordination. With respect to *Value*, we find that including an opportunity to reprioritize among shared goals (comparing Conditions 2 and 1) generates greater *Value*, 0.473 points higher on a five-point Likert scale ($p < 0.01$), representing a 0.786 standard deviation increase. The second mechanism of additional interim deadlines imposed by an additional meeting (comparing Conditions 3 and 2) generates greater *Value* than in Condition 2 by 0.295 points on a five-point Likert scale ($p < 0.05$), a 0.490 standard deviation increase. With respect to *Novelty*, we find an effect in the opposite direction. The addition of questions to reprioritize among goals (comparing Conditions 1 and 2) generates lower *Novelty* by 0.342 points on a five-point Likert scale ($p < 0.10$), a 0.494 standard deviation decrease. Increasing the frequency of meetings in Condition 3 generates a lower degree of *Novelty* than in Condition 2 with only one meeting, 0.397 points on a five-point Likert scale ($p < 0.05$) or a 0.574 standard deviation decrease.

Process

With respect to the process of integration, measured by *Time to Integrate*, we find that including an opportunity to reprioritize goals in Condition 2 (versus Condition 1) and the higher frequency of meetings in Condition 3 (versus Condition 2) lead to faster *Time to Integrate* ($p < 0.05$ and $p < 0.05$, respectively), suggesting that the separate components of iterative coordination do lead to integration. The effect of the additional goal question (Condition 2 vs. Condition 1) amounts to 396 seconds (6.6 minutes) faster *Time to Integration*, representing a 0.60 standard deviation reduction in how long it takes a team to use the whiteboard to integrate ideas, amounting to 11% of the total time (one hour) the team had available for the task. The effect of the additional meeting (Condition 3 vs. Condition 2) amounts to 376 seconds (6.3 minutes) faster *Time to Integration*, a 0.57 standard deviation reduction amounting to 10% of the total available time.

To evaluate specialization, we then use our data on the sketches generated by the individuals on each team as an indication of intermediate individual-level specialization activity. We find that both the goal reprioritization question in Condition 2 (versus Condition 1) and the additional meeting in Condition 3 (versus Condition 2) reduce the total count of *Individual Sketches* generated by the members of each team ($p < 0.05$ and $p < 0.01$, respectively). This result implies that the opportunity to reprioritize shared goals or increasing the frequency of meetings reduces the capacity of the team to be productive in preliminary individual specialization activity. The effect of the additional goal question (Condition 2 vs. Condition 1) amounts to 1.1 fewer pages of sketches, a 0.69 standard deviation reduction in sketches or lost sketch productivity of about 0.61 person-hour. The effect of the additional intervention (Condition 3 vs. Condition 2) amounts to 1.2 fewer pages of sketches, a 0.76 standard deviation reduction in sketches or lost sketch productivity of about 0.86 person-hour.[67]

[67] We calculate person-hour productivity in Condition 1 as the 1.8 pages of sketches generated by the average participant in an hour. In Condition 2, person-hour productivity is 1.4 pages of sketches per hour.

Additional Findings

We collect data for a number of supplemental analyses to shed additional light on the effects of different implementations of iterative coordination. First, we assess the impact on completeness to assess (and then rule out) the possibility that there is a general productivity effect, e.g., a world where iterative coordination positively (or negatively) impacts both value and novelty just because organizations get more (or less) work done on a general basis.

Second, we use the survey measures of coordination and specialization as further validation, from the participant perspective, of the processes that our experimental manipulation of iterative coordination engenders: increasing coordination for integration and reducing specialization. We confirm—via a post-experiment survey—that more thorough implementations of iterative coordination positively associate with self-reported measures of coordination and negatively associates with self-reported measures of specialization.

Third, we assess the amount of time iterative coordination takes because the time cost of iterative coordination is an important boundary condition for organizations deciding whether to adopt it. It is ex ante plausible that the time taken by the meetings would be meaningfully large and would hurt productivity on a general basis. We use the video recordings to code both the duration of iterative coordination meetings and the time it takes teams to get back to work after a meeting. We find that the meetings and time afterwards take up a small amount of time (3.6% of total available time), and adding a second meeting (Condition 3 vs. Condition 2) takes up less than double the time of the first meeting alone.

Fourth, we want to evaluate the communication taking place between and out of the iterative coordination meetings to get a sense of whether iterative coordination was increasing (complementary with) or decreasing (substituting for) the normal levels of communication that would otherwise take place if the organization did not practice iterative coordination. We measure the oral communication that takes place in between iterative coordination meetings using the video recordings. We find that adding the additional question (Condition 2 vs. Condition 1) and adding the additional meeting (Condition 3 vs. Condition 2) lead to

a statistically significant increase in the frequency of communication between meetings but not in the total amount of words being spoken.

Fifth, we analyze the oral communications of each team to build several text-based measures (Reypens and Levine 2018) to understand whether there is an effect on the salience of the final deadline (at the end of the sixty-minute period) and its implications for integration and affect, i.e., negative emotion or anxiety. We find that adding the additional question (Condition 2 vs. Condition 1) and adding the additional meeting (Condition 3 vs. Condition 2) lead to a statistically significant increase in the use of words associated with time, as well as a shift towards words proposing new activity distinct from prior activity. When combined with the other evidence, we interpret this to be a shift towards more integration. However, we do not find any evidence that a more thorough implementation of iterative coordination leads to more anxiety, negative emotion, or swearing as observable in oral communication. The Online Appendix reports the data collection methods and results for these analyses.

Together, these results suggest robustness to a number of alternate mechanisms. For instance, the time cost of (additional) meetings cannot account for the entire observed decrease in specialization among iteratively coordinating organizations. Although an additional meeting mechanically reduces raw time for work, the incremental time cost of a meeting decreases, i.e., a subsequent additional meeting takes less time than a previous meeting. Similarly, while additional meetings increase raw latency for teams to resume their work, this latency diminishes with increasing frequency of meetings.

4.4.4 Discussion of Follow-on Study

The laboratory experiment shows that more thorough implementations of iterative coordination have a positive association with value and a negative association with novelty. These findings align with the findings of the primary field study, which only compares iterative coordination against a baseline control condition with no intervention. Furthermore, more thorough implementations of iterative coordination associate with integrative activity—reflected in this study as the quicker integration of sketch material into the final product—while be-

ing negatively associated with individually specialized activity—the pages of individual draft sketches produced.

When testing two mechanisms of iterative coordination—namely, the frequency of interim deadlines and goal reprioritization questions—we find that augmenting these aspects of iterative coordination amplifies its effects. That is, the addition of discussion questions (creating opportunities for reprioritizing goals) and the additional meetings (imposing additional interim deadlines) yield stronger positive effects on value and stronger negative effects on novelty. These two mechanisms similarly amplify the positive effect on integration and the negative effect on specialization associated with iterative coordination in general. In particular, additional deadlines lead to a shift in activity towards integrating knowledge. This shift is triggered by an attention to time, a focus on the future deadline, and a recognition of discrepancy between the current state and the future desired state. Interestingly, additional interim deadlines from iterative coordination do not increase anxiety among participants. Here, we posit that the lack of anxiety may relate to an ability to reprioritize goals using iterative coordination. Prior work demonstrates how unfilled goals make individuals anxious (Masicampo and Baumeister 2011). By reprioritizing goals, individuals treated with iterative coordination may create new plans for unfilled goals, helping reduce their anxiety (Masicampo and Baumeister 2011). Overall, for business practice, the mechanisms of frequency of interim deadlines and goal reprioritization questions allow mangers and organizations to tweak iterative coordination to generate the degree of value and novelty they desire.

In interpreting the mechanisms underlying the questions, in Section 4.2 we particularly focus on Question 3's influence on helping the organization revisit its overall priorities and potentially adjust them in the future (by asking, for instance, if goals have changed). This idea of adjustment in prioritization over time is central to Agile practice more generally (Sutherland and Sutherland 2014, Rigby et al. 2016b). Theoretically, it permits satisficing behavior where relaxing originally stated objectives allows for more attainable standards on

objectives—that may fail to be met due to ongoing underperformance (Simon 1947, Hu and Bettis 2018)—such as the acceptable level of novelty in the final product. Future work could examine whether ex ante specifying the goals to be discussed in the third discussion question to explicitly include novelty may help preserve outcomes of novelty over time, or if it would instead lead to underperformance on both novelty and value objectives.

4.5 Discussion and Conclusion

How organizations pursue multiple performance goals for which there is no clear prioritization among them is an important question for organizations research. In this paper, we study the effects of the widespread yet poorly understood practice of iterative coordination, which allows organizations to prioritize among novelty and value in innovation. We embed a field experiment within a software development competition in partnership with Google. We find that while iteratively coordinating firms develop products that are more valuable, these products are simultaneously less novel. By tracking the underlying software code, we find that iteratively coordinating firms integrate existing knowledge at the cost of specializing to create new knowledge. We then conduct a follow-on laboratory experiment to help verify underlying mechanisms for iterative coordination. We find that increasing the frequency of meetings and including an opportunity to reprioritize organizational goals amplify integration in iterative coordination, driving value at the cost of novelty in outcomes.

We now detail three primary contributions of this work: highlighting how iterative process to manage multiple performance goals may *implicitly* prioritize certain outcomes in innovation; introducing software code tracking methodology for studying organizational innovation process; and addressing recent calls from the literature to study new methods of organizing, especially in emergent contexts such as hackathons.

4.5.1 Prioritizing Multiple Goals and Outcomes in Innovation

Our findings suggest that iterative coordination causes an organization to *implicitly* prioritize value over novelty in innovation. This prioritization is implicit: even when an organization believes in preserving novelty in outcomes, the process of iterative coordination may drive it

to ultimately favor value. With additional deadlines, iterative coordination drives integration at the cost of specialization—processes that we document in the software code from the primary field experiment and in the product development activities in the follow-on laboratory experiment. These processes, in turn, associate with outcomes of value over novelty. This finding stands out because iterative coordination itself does not explicitly define a priority for either novelty or value.

Although the literature on managing multiple performance goals examines the influence of frequent and repeated coordination on performance broadly defined (Gavetti et al. 2012, Knudsen and Srikanth 2014, Levinthal and Workiewicz 2018), coordination's influence on novelty and value in outcomes remains underexplored. To contextualize our findings with respect to prior work, we can consider the literature on goal-setting in creativity, which predicts that frequent and repeated coordination on innovation goals yields both greater novelty and value in outcomes (Carson and Carson 1993, Byron and Khazanchi 2012). The assumption is that the process of revisiting and iterating in innovation-oriented goals reinforces individual motivation, an important antecedent to innovative outcomes (Cromwell et al. 2018b). Especially for more dynamic organizational activity such as new product development, a process for iterating on goals in innovation is argued to help keep organizational members aligned in their pursuit of innovation in the face of failures (Alexander and Van Knippenberg 2014). Because failures in the pursuit of innovation can have a demotivating effect, scholars have argued in favor of using frequent and repeated coordination to help promote novelty and value in end outcomes (e.g., Amabile and Pratt 2016).

Given the theorized benefits of goal-setting-driven motivation for producing novelty and value in end outcomes, we could also expect these benefits to manifest in other dimensions of innovation performance. However, our empirical evidence suggests no differences in general productivity from our coordination interventions in innovation projects, both in terms of aggregate code contributions in the primary field study and completeness of the final task in the follow-on laboratory study. Although we do not directly measure effects on moti-

vation, it is theoretically unclear from prior work whether goal-setting's motivational benefits should manifest themselves in terms of greater novelty *and* value—the two constituent parts of innovation—or if the effect is channeled into just one of these two outcomes, such as value. In other words, it is not clear from prior literature whether we should expect motivation's effects to improve innovation across end outcomes as a whole, or whether it more narrowly drives parts of innovation, such as value.

In contrast to the predictions of the literature on goal-setting in creativity, our findings suggest that frequent and repeated coordination on innovation goals leads to improved value, but to *reduced* novelty. This insight derives in part from our conceptualization of the innovation process as the pursuit of the *distinct* goals of novelty and value. Prior work often conceptualizes creativity and innovation as a joint outcome of novelty and value (Oldham and Cummings 1996, Shalley and Perry-Smith 2001). Accordingly, the joint view inherently underemphasizes the costs associated with coordinating individuals on potentially competing goals of novelty and value. Knudsen and Srikanth (2014) describe coordination costs as those associated with communicating among multiple organizing members, which inherently takes time away from specialist work. Our analysis suggests that while greater coordination on innovation goals may help drive integration and the pursuit of value, it does so at the cost of specialization and the pursuit of novelty as a *distinct* goal. As a result, even if organizational members feel they can choose to pursue either novelty or value, we argue that iterative coordination and related management practices drive an iterative process that ultimately results in the implicit prioritization of value over novelty. This reprioritization towards value over novelty is especially salient in supplemental empirical findings from the primary study that distinguish originally stated objectives from what gets developed in final projects.[68]

Having established the importance of studying innovation as distinct goals of value and novelty, our findings also contribute to the behavioral theory of the firm (BTF) by

[68] Online Appendix C.2 presents this supplemental analysis.

explaining how satisficing occurs across multiple, potentially conflicting goals in innovation. A long stream of research on the BTF long posited that decision-makers respond to search across multiple performance objectives by satisficing (Simon 1955, Posen et al. 2018), or choosing the first alternative they expect to satisfy all objectives (Gavetti et al. 2012). While this literature explores satisficing behavior with respect to a single objective (Posen et al. 2018), exactly how organizations satisfice across multiple objectives remains unclear, with Greve and Gaba (2020, p. 323) noting that "the theoretical treatment of adaptive behavior amid multiple goals and performance aspirations has only begun." This research has been limited in part due to the ambiguity of performance feedback when simultaneously pursuing multiple objectives. For example, it is unclear how organizations respond when an alternative is judged to be a success on some objectives but a failure on others (Levinthal and Rerup 2020). Given mixed signals of success and failure across multiple objectives, on which objective does the organization focus? In early work on this domain, Gaba and Greve (2019, p. 647) suggest that "the goal perceived as more important for survival gets priority and triggers stronger reactions." However, in the pursuit of innovation, it is unclear whether value or novelty would or should be prioritized to ensure organizational survival. Does an organization embrace outcomes of high value that would be of known interest to customers, or does it instead focus on novelty to distinguish itself among customers? This fundamental ambiguity exists for both innovating organizations and the scholars that seek to study those organizations.

Our findings suggest that satisficing in innovation occurs specifically in favor of value and against novelty. What is particularly striking about our findings is that regardless of an organization's starting point in the balance between novelty and value, the organization deprioritizes novelty. When interpreting feedback across multiple objectives, Levinthal and Rerup (2020) suggest that a process of self-enhancement may occur, where organizations prioritize the dimensions on which they are performing the strongest. Such a response follows from the intuition that the very first alternative that is deemed satisfactory is ultimately

chosen (Gavetti et al. 2012). In other words, if an organization is already performing relatively better on novelty than value, why not focus on novelty? A promising area of future inquiry would be to understand exactly why organizations do not adopt a self-enhancement perspective to emphasize novelty under frequent and repeated goal coordination.

4.5.2 Methodological Contributions

As a second primary contribution, we are among the first researchers in this literature, to the best of our knowledge, to measure organizational innovation in real time using version-control software. Tracking software code development provides significant advantages towards studying innovation across individuals, firms, and time, which we detail below.

First, it allows for the measurement of productive and creative tasks *across individuals*. In particular, extant literature in strategy for how organizations generate innovation commonly brings up the concepts of specialization and integration as processes by which innovation might emerge (Ethiraj and Levinthal 2004, Levinthal and Workiewicz 2018). However, prior empirical work, e.g., using patent data or other final outputs, often cannot directly measure how much individuals in the organization work together and integrate their ideas across individuals in real time. We can track in the software code development process when code is being integrated across individuals (i.e., *Code Integration Action*). There are plenty of opportunities for future studies to use code tracking of across-individual activity, e.g., integration across managers and subordinates (e.g., Reitzig and Maciejovsky 2015, Ghosh et al. 2020), division of labor in innovation production (e.g., Lawrence and Lorsch 1967, Knudsen and Srikanth 2014, Ghosh 2020), and the extent to which ideas are combined with one another or discarded from consideration (e.g., Girotra et al. 2010).

Second, it allows for the measurement of innovative activity *across firms*. Our measure of *Advanced API Specialization* reflects whether firms put effort into learning and utilizing sophisticated developer tools offered by Google. Although we use this measure in a relatively limited way, this measure can be interpreted more broadly in the context of a rapidly growing literature on multi-sided platforms (Wen and Zhu 2019, Pan Fang et al. 2020). This set

of developer tools can be thought of as a multi-sided platform, where our firms are the complementors to the Google platform (and consumers are the other side of the platform). Future studies can use this type of data to identify when and how firms draw on external platforms or tools for the development of innovation. For example, this data could be used to measure precisely the knowledge-driven factors leading to phenomena such as how entrepreneurs draw on knowledge from larger organizations, particularly as they depart from a larger organization to launch their own firm (Kacperczyk and Marx 2016).

Third, it allows for the measurement of individual innovative activity *across time*. Despite the common notion that innovation might arise in a flash of genius, scholars recognize that innovation is often an emergent process that occurs over time (Amabile and Pratt 2016, Hu and Bettis 2018). In our experimental setting, we can allow for our developers to innovate over the course of a day, but in most settings this innovative process runs over the course of months or years. By longitudinally tracking software code across time, we can capture with greater granularity how innovation arises. In particular, we show that iterative coordination has an additive or cumulative effect of pushing firms more towards integration activity and away from individually specialized activity. Future work can use this methodology perhaps even on a much grander time scale to look at how other organizational characteristics such as structure or individual characteristics like managerial cognition change the patterns of the innovation process over time (Schilke et al. 2018).

4.5.3 New Methods of Organizing and Hackathons

Finally, we respond to recent calls from the literature to study new forms in organizing (Levine and Prietula 2014, Puranam et al. 2014, Burton et al. 2017), particularly in the context of innovation at hackathons. Over the last several years, hackathons have emerged as an important context for organizational innovation (Pan Fang et al. 2020, Pe-Than et al. 2019), with startups and incumbents alike leveraging the context to develop ideas and projects of high value and novelty. Lifshitz-Assaf et al. (2020, p. 4) note that since "hackathons adhere to non-hierarchical and open ways of organizing, no clear process, structure or roles [are] de-

fined." In such an environment, organizations need new structures and practices to sustain high performance (Meyer and Zucker 1989). In the absence of traditional process, structure, or roles in hackathons, new organizing practices like iterative coordination may be necessary to manage innovation in the highly time-constrained setting of a hackathon.

Where prior work finds that coordination frameworks from contexts brought in from outside a hackathon may be detrimental to innovation (Lifshitz-Assaf et al. 2020), our findings suggest that the signature Agile practice of iterative coordination can be effectively employed to produce complete innovation projects at hackathons. In contrast to prior findings suggesting that rapid coordination efforts in a hackathon setting may lead to project failure (Lifshitz-Assaf et al. 2020), we find that the high degree of coordination engendered by iterative coordination does yield complete, functioning output by the end of a hackathon. It is important to reflect on why our findings differ. Whereas Lifshitz-Assaf et al. (2020) suggest that coordination frameworks may cause firms to bound themselves to ultimately unattainable goals of both high value and high novelty, our findings demonstrate that iterative coordination allows firms to reprioritize their goals over time, allowing critical adjustments to be made regarding the degree of novelty that is acceptable to them over the course of their projects.

Of course, it is important to note that iterative coordination does not necessarily yield products that are judged to be more complete than a baseline of minimal coordination; rather, iterative coordination yields successfully completed projects which bias towards value instead of novelty. This raises the question: What may motivate a firm to use iterative coordination to manage innovation projects at a hackathon? One factor may be to appropriately balance value against novelty in nascent projects. For instance, in corporate hackathons, projects that are judged to be too new and that are difficult to relate to a focal firm's business value objectives are often abandoned after a hackathon and are not pursued for further development (Nolte et al. 2018). In this sense, an idea or proposal that is judged to be too new while offering little value has little chance of being implemented by the firm

(Ahuja and Morris Lampert 2001). Although traditional coordination methods may stifle projects at a hackathon, iterative coordination could allow firms to inject necessary value into ideas, increasing their likelihood of being implemented.

4.5.4 Limitations and Future Work

We conclude by noting limitations to the present study and opportunities for future work. First, we note that while many contexts in which iterative coordination is used deeply prioritize innovation (Rigby et al. 2016a), iterative coordination itself does not directly frame a need for novelty or value. Given our interest in evaluating iterative coordination as it is practiced, we intentionally do not frame any of the three iterative coordination questions with an additional requirement that the organizational goal (or its subsequent output) be novel or valuable. Varying the composition of iterative coordination questions to articulate an explicit need for novelty, value, or other outcomes is a promising area for future study. In addition, given our empirical focus on entrepreneurial innovative settings, our theoretical treatment of iterative coordination assumes its application in settings in which existing technical knowledge is comparatively limited; examining the applicability of our theory to settings in which technical knowledge is well established would be a worthy topic of future inquiry.

There are several potential organizational moderators to iterative coordination left to be explored. The distribution of the quantity or quality of individual specialization—whether individual contributions are evenly divided or concentrated in one individual—might affect the degree of integration that would occur when practicing iterative coordination. The size of the practicing organization may also moderate the effects of iterative coordination, e.g., large organizations have a greater need for coordination (Aggarwal et al. 2020), but at the same time iterative coordination may not scale well to a large number of participants in a meeting. In addition, characteristics of the goals themselves, such as being too specific or too challenging, may potentially moderate iterative coordination's influence on innovation (Ordóñez et al. 2009). Given limitations on statistical power in the present study, we

recommend the exploration of these moderators for future study.

Finally, across practical contexts, there may be heterogeneity in the effects of iterative coordination. For instance, the design of physical products and/or services, as opposed to software development, may have fundamentally different organizing needs. This may arise due to alternate environments of complexity and modularity inherent in the architecture of the offering (Ulrich 1995). Of course, differences in underlying complexity and modularity may call for different methods of coordination in innovation (Baldwin and Clark 2000, Ethiraj and Levinthal 2004).

Appendix A

Appendix to Think Before You Act: The Unintended Consequences of Inexpensive Business Experimentation

A.1 Appendix: Institutional Context

A.1.1 Code Change Examples

Treatments that require more cognitive resources require back-end code customization on behalf of organizations. Consider the example of a financial services firm considering how to increase conversion as measured by monthly retirement account contributions. An example of a complex, interdependent hypothesis to test may be that "peer-effect information within one's age cohort will increase retirement contributions" (Beshears et al. 2015). Retrieving this peer-effect information and accurately displaying it to users requires significant back-end code customization on behalf of the organization, which must retrieve the user's age and match it to information from the company's databases.

The aforementioned treatment cannot be successfully created via Optimzely's standard "What You See is What You Get" (WYSIWYG) development environment. This visual editor is best suited for simple, static changes to a website, such as changes to colors, headlines, and the positioning of elements. A comparison of the visual editor for WYSIWYG edits and the JavaScript custom code editor is displayed in Figure A.1.

— Insert Figure A.1 About Here. —

A.2 Appendix: Data and Measures

A.2.1 Data Construction

A basic requirement for a business experiment is that it generates information that helps the organization learn (Thomke 2003, p. 98) and that is relevant to a firm's strategy or configuration of choices (Rivkin 2000, Ghosh et al. 2020). Not all tests on Optimizely's platform qualify as true experiments. For instance, experimenters may conduct "A/A" tests, where the baseline is compared to itself in order to measure the quality of experimentation infrastructure. Mistakenly including A/A tests in our sample would bias our estimates of cognition's impact on performance, as A/A tests do not test any changes from which an organization may learn.

In addition, it is crucial to define when an experiment ends. While most tests run only as long as necessary to establish statistical power, in practice some run for longer than necessary. Without a natural stopping rule for experimental search where at least two paths are tested and one is ultimately chosen (Gans et al. 2019), the A and B variants of an A/B test become real options for an organization. Business experiments differ from real options in their intent to promote learning, where in contrast to real options, experiments "not only provide information about intended investment paths but also provide information about other possibilities—possibilities that may not even have been envisioned at the time of initial investments" (Adner and Levinthal 2004, p. 77). As real options are not structured to promote the sequential learning and adjustment that is key to experimentation (Thomke and Bell 2001), they are not business experiments (Contigiani and Levinthal 2019).

Other examples of actions taken on the Optimizely platform that are not business experiments include A/A tests, "hotfix" behavior to quickly implement changes and bypass formal release cycles, and tests that do not include a treatment variant tested against a control.

A.3 Appendix: Interdependent Change and Performance

To probe the robustness of my findings on the association between interdependent change and performance in A/B testing, I conduct a series of stress tests that are each individually detailed below.

A.3.1 Interdependent Change and Alternate Performance Measures

Although my primary dependent variables focus on performance breakthrough and failure, it is important to assess the robustness of my findings to measurements of mean lift. Here, I measure *Lift*, which is the percent improvement in conversion rate due to experimental treatment. In addition, I also measure *Statsig Lift*, which measures lift conditional on an experimental treatment achieving statistical significance. While my interviews reveal that many practitioners are comfortable making decisions on the basis of raw lift results (unconditioned on statistical significance), others are only comfortable interpreting the performance of a treatment should the effect by deemed significance by Optimizely X's stats engine (for further detail, see Pekelis et al. (2015)).

Associations between *Code Change* and mean lift outcomes are displayed in Table A.1. My findings are consistent across models, including those with the full set of controls. Note that the observation count decreases in models running *Statsig Lift* due to censoring on the basis of whether statistical significance was achieved.

— Insert Table A.1 About Here. —

A.3.2 Alternate Measures of Code Change

While my interviews demonstrate that lines of JavaScript code are a reliable measure of cognitive investment into a treatment (especially as more complex, interdependent resources draw upon data resources within a website's back-end, see Section A.1 for additional detail), there are potential concerns regarding heterogeneity in coding style. While engineering development resources remain relatively steady within organizations over time (suggesting that across firm variation in coding style will be captured by organizational fixed effects), I construct an alternate measure for code change which counts the number of characters written

in JavaScript code. This alternate measure may, for instance, more directly capture total code contributions, including script line length and the comments which reflect additional cognitive effort. I test this alternate measure in Table A.2 and find consistent results with Tables 2.3 and 2.4.

— Insert Table A.2 About Here. —

A.3.3 Robustness to Low Baseline Conversion

Lift offers a number of attractive features for performance measurement, such as the built-in normalization of effects which enables comparisons across experiments in terms of relative increases (Gordon et al. 2019). However, its weakness stems from the potential to overstate effects when considering low baseline conversion rates, which are the denominator in the calculation of lift.

To ensure that my results are not solely driven by experiments run on websites for which there were relatively low baseline conversion rates, I rerun my analyses on the sample of experiments for which baseline conversion rates exceeded 0.1%. This threshold was determined from conversations with Optimizely engineers and data scientists about data quality, but my findings are not sensitive to this particular cut-off. The results of this robustness check are reported in Table A.3.

— Insert Table A.3 About Here. —

A.4 Appendix: Action Resources and Interdependent Change

To probe the robustness of my findings from my main models measuring feedback resources' effects on action and cognition, I conduct a series of stress tests that are each individually detailed below.

A.4.1 Robustness to Count Models

Given that my primary dependent variables in Section 2.5 are count variables, I assess the robustness of my results using conditional logit fixed models. While the Poisson model is well-suited for count variables, it traditionally suffers from an incidental parameters problem in the face of fixed effects, which are core to my empirical strategy. Given the potential for

an inconsistent estimator and harder to interpret effects, I elect to use OLS fixed effects models in my main analysis.

Nonetheless, I assess the robustness of my specification to a Poisson model in Table A.4 and find consistent results across models with and without traffic and experimentation program controls.

— Insert Table A.4 About Here. —

A.4.2 Alternate Measure of Traffic

As described in Section 2.5, a critical resource for front-end conversion-rate A/B testing is top of the funnel website traffic. This is typically observed for websites at a main domain level and is recorded in terms of total page views. A potential weakness to measuring total page views is that it may not reflect the number of total unique visitors who interact with a website during a given month. That is, total page views may be driven up by internet users who repeatedly visit a website. This is in contrast to truly unique leads that many argue form a fundamental basis for website growth and performance. As a result, when making resource allocation decisions, some site administrators prefer to track *Unique Visitors*, which measures the number of unique website visitors over the course of a month. I collect this data from the SimilarWeb Desktop Traffic API and merge it with my two-year panel on organizational experimentation.

In Table A.5 I reproduce analyses from Table 2.6 but with a new measure for traffic using *Unique Visitors*. My findings are consistent to those of my analyses on the main paper, where increases in *Unique Visitors* associate with additional experiments but reduced cognitive investment per experiment. Note that the reduced observation count reflects the lack of availability of *Unique Visitors* data from SimilarWeb for a subset of organizations in the full Optimizely sample.

— Insert Table A.5 About Here. —

A.5 Appendix: Google Natural Experiment

A.5.1 Institutional Context

Google LLC is a leading provider for inbound traffic for web enterprises today. Given Google's pivotal role in helping web enterprises secure traffic at the top of the conversion funnel, many enterprises devote considerable attention to search engine optimization efforts intended to help their enterprises rise in the rankings of search results (Eric et al. 2015). Nonetheless, Google's value proposition relies on its ability to accurately serve the user with the most relevant and important content related to their search query first. To prevent web enterprise from gaming Google's rankings of websites, Google remains secretive about the inner working of its algorithms.

To improve its service, Google periodically makes updates to its search algorithm and the layout of the "search engine results page" (SERP), which is the page that is displayed to the user after entering a Google search. These improvements are made secretively and without notice to the world's websites. While focused on improving the experience of their users, Google's SERP changes often create unintended winners and losers. That is, some websites find gains in traffic, while others often suffer considerable losses in traffic from Google. Crucially, these SERP changes and their subsequent effects are unanticipated by websites by design.

While algorithm changes are often made quietly and unbeknownst to Google's users, in June 2018, Google instituted a major SERP change by introducing a video carousel in the first page of results. This change was intended to mirror a mobile browsing experience, where video content that is relevant to a user's search query is served up front. In practice, this change ended up moving certain organic blue links into the video carousel.

While video content may seem appealing to users, in practice, many search queries suffered from the change. In particular, many online retailers suffered considerable losses in source traffic from Google. Reflecting on the effects of the SERP change, an anonymous Optimizely client in retail offered the following:

> "Our [marketing] folks were screaming murder... The working theory was that leads would hover right past those video boxes and click on other lower content [down the first page of results]."

This testimonial was confirmed in case studies released shortly after the change—where potential customers were not conditioned to find relevant content in the video carousel (Gabe 2018). As a result, they would scroll past these results, which counterfactually existed as organic blue links on the SERP. A screenshot from a sample query is provided in Figure A.2.

Given the unanticipated and quasi-exogenous nature of this change, I use it to identify the effects of a decrease in the resources for action in experimentation. Below, I describe two methods to estimate the effects of this quasi-experiment: first, in a difference-in-differences event study on the implementation of the SERP change on experimentation in retail, and second, via an instrument that estimates the effect of Google source traffic on overall website traffic.

— Insert Figure A.2 About Here. —

A.5.2 Retail Difference-in-Differences Estimation

Data and Variables

Using the organization-month panel and outcome variables described in Section 2.5.1, I introduce a dummy variable *Retail*, which takes a value of 1 when the organization is classified by Optimizely as being an online retailer.

Estimation Strategy

I estimate the effects of the Google SERP change on retailers in the following differences-in-differences model:

$$Y_{it} = \beta(Retail_i \times Post_t) + \alpha_i + \delta_t + X_{it}B + \epsilon_{it}.$$

Y_{it} represents the dependent variables of *Experiment Count* and *Mean Code Change*. $Retail_i$ is an indicator variable taking a value of 1 for retail organizations, and $Post_t$ is an

indicator variable equaling 1 for the months during and after the Google SERP change. Our coefficient of interest is β, which estimates the effect the Google SERP change on Y_{it}. X_{it} is a vector of time-variant organizational controls, α_i is an organization fixed effect that controls for time-invariant unobserved confounding factors, and δ_t is a month fixed effect to control for potential shocks across all organizations during the observation period.

A.5.3 Google Traffic IV Estimation

To more directly estimate the effects of the Google SERP change, I propose an instrumental variables method which uses Google source traffic as an instrument on the overall traffic received by an organization.

IV Assumptions

Crucial to an instrumental variable strategy are the strong first-stage and exclusion restriction assumptions. While the first assumption is testable (the results of which are displayed in Table 2.9), the latter is not.

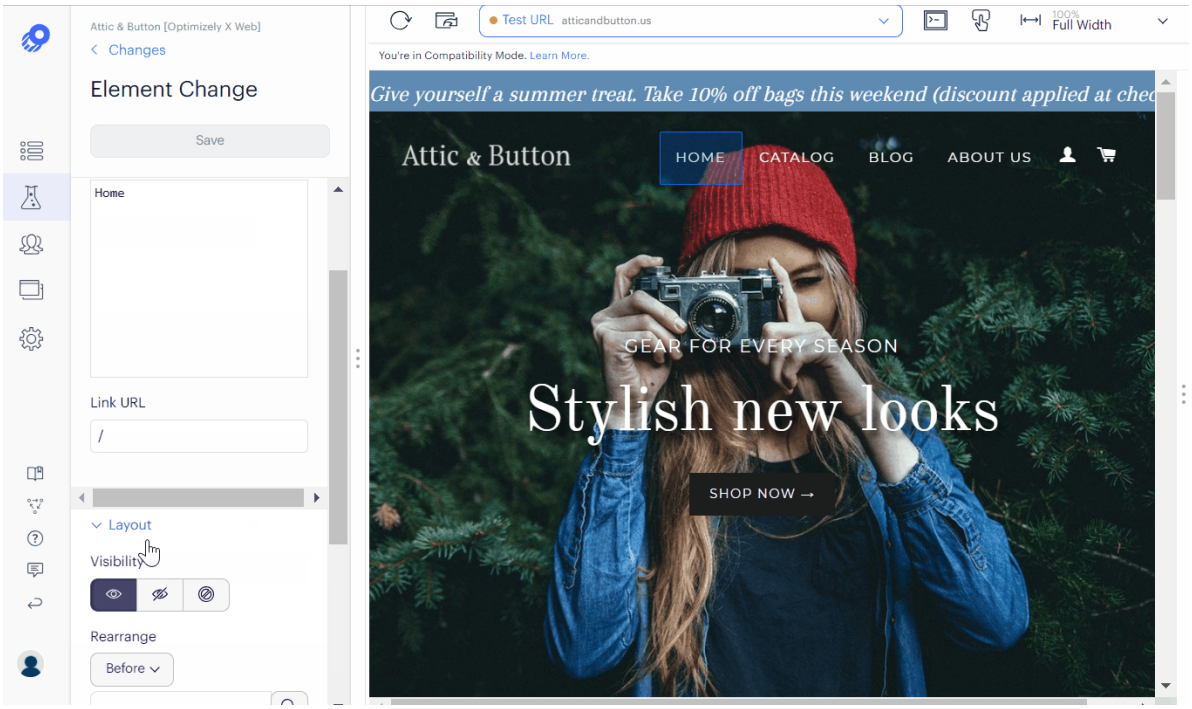
The exclusion restriction states that the effects of the Google SERP change can only act through the channel of overall website traffic. There are several contextual factors that support this assumption. First, my sample is restricted to bottom of the funnel tests, helping us avoid top of the funnel, SEO tests dedicated specifically towards increasing inbound traffic. Second, my interviews with Optimizely clients indicate that experimenters overwhelmingly make resource allocation decisions for experiments on the basis of traffic that reaches the bottom of the funnel, where customer conversions occur. Finally, the use of Optimizely A/B testing to trick Google search crawlers into believing that a page should receive a higher rank is explicitly forbidden. In particular, the practice of "cloaking," where a Google search crawler is shown a different site (i.e., a "B" variant) than the original site (i.e., an "A" variant), is forbidden by Google's terms of service. Google reserves the right to remove sites from its search results that it believes are cloaking. As a result, Optimizely users are actively discouraged from testing in a way that directly targets increases in Google source traffic.

Data and Variables

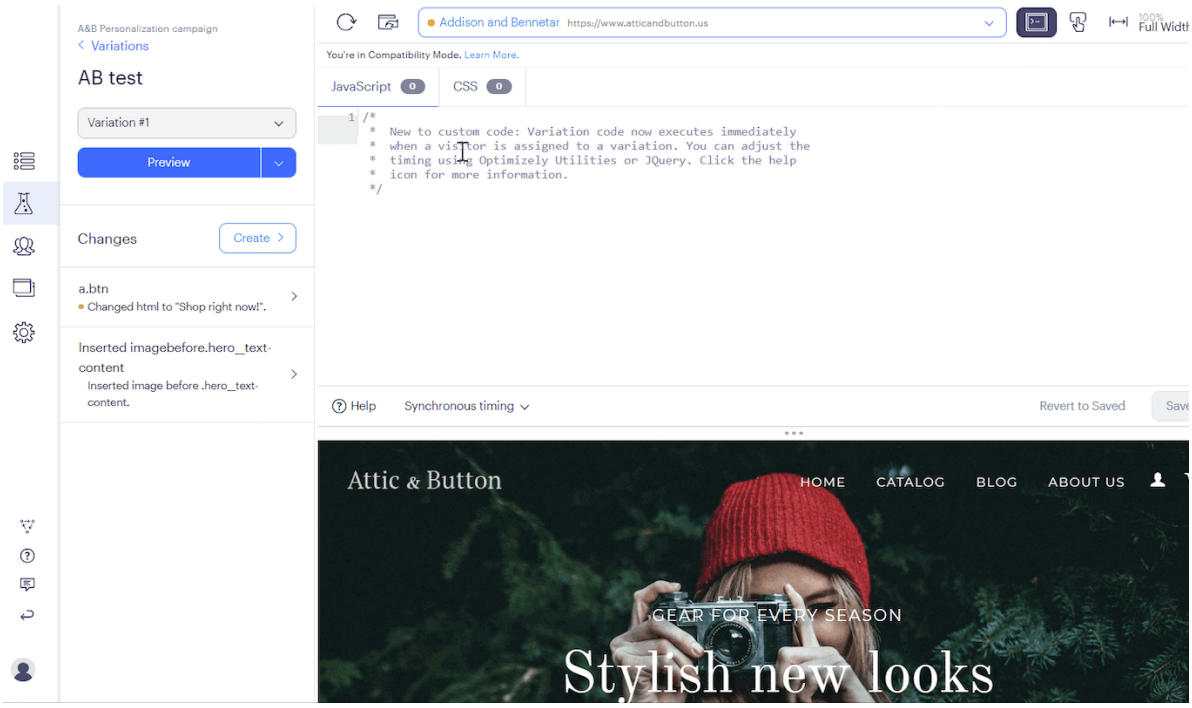
For the Google instrument, I collect data from the SimilarWeb's web traffic source APIs. I construct the variable *Google Traffic*, which is the logged number of pageviews sourced from Google organic search in a given month.

Estimation Strategy

I instrument for *Traffic* using *Google Traffic* in a two-stage least squares (2SLS) model with organization and month fixed effects. Robust standard errors are clustered at the organizational level.

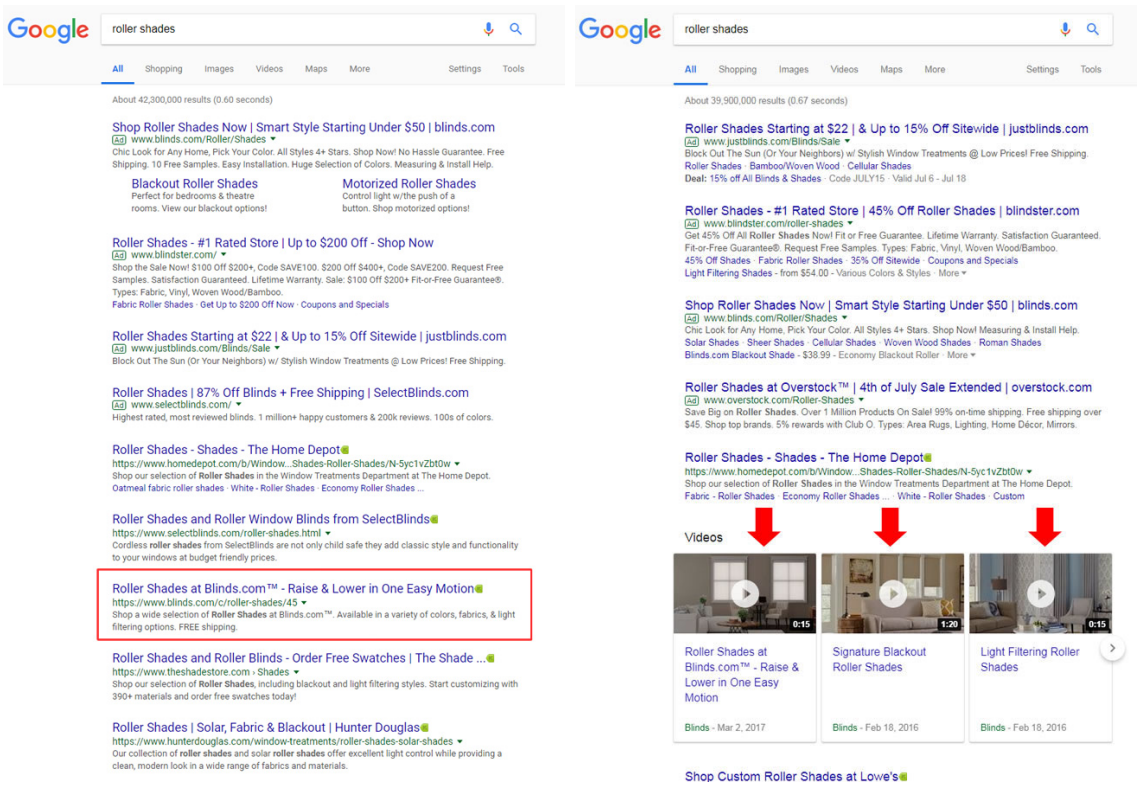


(a) Visual Editor



(b) Custom Code Editor

Figure A.1: **Visual vs. Custom Code Editor.** Panel (a) displays the visual editor mode of the Optimizely platform, where users can make point-and-click edits to their websites. In this example, the experimenter is making edits to the layout of the "Home" button on the webpage. Panel (b) displays the custom code editing mode, intended for high-interaction changes to websites.



(a) Pre-Carousel SERP (b) Post-Carousel SERP

Figure A.2: **Google SERP Natural Experiment**. A comparison of the Google search engine results page for a search of "roller shades." Here, the organic listing circled in Panel (a) is replaced by video carousel links in Panel (b). This change led to traffic losses for the Retail firms in the Optimizely sample. Screenshots of the example search are reproduced from Gabe (2018).

Table A.1: **Alternate Performance Measures: Mean Lift.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Lift‡		Statsig Lift‡	
	(1)	(2)	(3)	(4)
Code Change‡	0.02801****	0.02552****	0.03795****	0.03819****
	(0.00646)	(0.00644)	(0.01042)	(0.01065)
Duration		0.00346**		0.00367*
		(0.00169)		(0.00189)
Sample Size‡		0.00408		0.00647
		(0.00504)		(0.00795)
Variant Count		0.06567****		0.02030**
		(0.00910)		(0.00960)
Metric Count		0.00053		0.00060
		(0.00177)		(0.00248)
Development Time		0.00025		-0.00093*
		(0.00026)		(0.00052)
Prior Experiments		-0.00035		0.00031
		(0.00029)		(0.00059)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Metric FE	Yes	Yes	Yes	Yes
Observations	31,716	31,716	6,233	6,233

Table A.2: **Alternate Measurement of Code Change: Characters Changed.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Top 5% Lift		Bottom 5% Lift	
	(1)	(2)	(3)	(4)
Change Characters‡	0.00218****	0.00197***	-0.00172***	-0.00156***
	(0.00064)	(0.00064)	(0.00054)	(0.00056)
Duration		0.00046		0.00015
		(0.00033)		(0.00037)
Sample Size‡		-0.00228*		-0.00250**
		(0.00121)		(0.00103)
Variant Count		0.00969****		-0.01387****
		(0.00219)		(0.00137)
Metric Count		0.00033		0.00004
		(0.00033)		(0.00030)
Development Time		0.00000		0.00006
		(0.00006)		(0.00007)
Prior Experiments		-0.00008**		-0.00003
		(0.00004)		(0.00004)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Metric FE	Yes	Yes	Yes	Yes
Observations	31,716	31,716	31,716	31,716

Table A.3: **Subset Analysis: Omitting Low Baseline Conversion Tests.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Top 5% Lift		Bottom 5% Lift	
	(1)	(2)	(3)	(4)
Code Change‡	0.00322**	0.00311**	-0.00282**	-0.00262**
	(0.00126)	(0.00126)	(0.00113)	(0.00115)
Duration		-0.00009		0.00003
		(0.00029)		(0.00041)
Sample Size‡		-0.00290**		-0.00307***
		(0.00113)		(0.00112)
Variant Count		0.01039****		-0.01456****
		(0.00214)		(0.00153)
Metric Count		0.00010		-0.00009
		(0.00031)		(0.00028)
Development Time		-0.00002		0.00008
		(0.00005)		(0.00007)
Prior Experiments		-0.00002		-0.00002
		(0.00003)		(0.00004)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Metric FE	Yes	Yes	Yes	Yes
Observations	28,743	28,743	28,743	28,743

Table A.4: **Alternate Specification: Poisson Model.** Variables demarcated by ‡ are natural log transformed. Robust standard errors are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Experiment Count		Mean Code Change	
	(1)	(2)	(3)	(4)
Traffic‡	0.232****	0.204***	-0.242***	-0.246***
	(0.068)	(0.063)	(0.094)	(0.093)
Pages Per Visit		0.007		0.026
		(0.008)		(0.017)
Direct Leads		-0.401		0.503
		(0.261)		(0.563)
EU share		-1.745****		0.121
		(0.485)		(0.823)
Infrastructure Testing		0.004		-0.023
		(0.005)		(0.024)
Underpowered Testing		0.044****		0.049****
		(0.011)		(0.010)
Metric Count		-0.012**		0.007
		(0.005)		(0.008)
Mean Development Time		0.003****		0.005****
		(0.000)		(0.001)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	37,200	37,200	26,688	26,688

Table A.5: **Alternate Measurement: Unique Visitors.** Ordinary least squares (OLS) estimation of organization-month level data. Variables demarcated by ‡ are natural log transformed. Robust standard errors clustered at the organizational level are shown in parentheses, with significance indicated by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Experiment Count‡		Mean Code Change‡	
	(1)	(2)	(3)	(4)
Unique Visitors‡	0.051**	0.054**	-0.095**	-0.098***
	(0.025)	(0.023)	(0.039)	(0.038)
Pages Per Visit		0.002		-0.007
		(0.003)		(0.004)
Direct Leads		0.093		-0.105
		(0.083)		(0.150)
EU share		-0.573****		-0.136
		(0.151)		(0.307)
Infrastructure Testing		0.026**		0.019
		(0.013)		(0.012)
Underpowered Testing		0.016****		0.028****
		(0.004)		(0.008)
Metric Count		-0.001		0.002
		(0.002)		(0.004)
Mean Development Time		0.003****		0.006****
		(0.000)		(0.001)
Organization FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Observations	31,920	31,920	31,920	31,920

Appendix B

Appendix to Do Senior Managers Help or Hurt Business Experiments? An Exploratory Study of Online Testing

B.1 Data and Measurement

Our primary measure of performance is *Max Lift*, which represents the maximum lift on the primary metric across n variants of an experiment. 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.

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. 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 $1 - \text{FDR} > 90\%$. Here, the use of 90% reflects standard industry practice of using a 10% threshold to deem statistical significance.

Table B.1: **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
	(1)	(2)	(3)	(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]
Organization 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 Integrations	0.002	0.003	-0.006	-0.012***
	(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 B.2: **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
	(1)	(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]
Organization 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 B.3: **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
	(1)	(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]
Organization 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

Appendix C

Appendix to Iterative Coordination and Innovation: Prioritizing Value over Novelty

C.1 Appendix: Design of Primary Study (Software Field Experiment)

C.1.1 Spatial Setup

Figures C.1 and C.2 depict the overall event space floor plan and detailed room floor plans, respectively. Participants in the control and treatment condition are in different rooms, so they never observe the mentor–participant interaction of participants in the other condition. There was limited opportunity for cross-condition social interaction during the competition; participants could interact prior to the competition, during registration and the welcome presentation, and after the competition during the dinner and the awards ceremony.

— Insert Figure C.1 About Here. —

— Insert Figure C.2 About Here. —

C.1.2 Recruitment Materials

Figure C.3 documents the solicitation content—an event website and an email template—used to recruit software developers to the hackathon field experiment. For email messages sent to industry professionals on behalf of our co-sponsor, this email message was used as a template. The co-sponsor could customize this message, although they may not promise any additional benefits or features of the competition beyond what is already described at the website.

— Insert Figure C.3 About Here. —

C.1.3 Competition Guidance

As part of the official guidance to participants in the competition, Google communicated the need for solutions that would be both novel and valuable to customers—thereby articulating the multiple goals of innovation. Figure C.4 depicts one of the slides used in the opening presentation by hackathon cosponsor Google to inform participating software developers of this need.

— Insert Figure C.4 About Here. —

C.1.4 Mentor-Participant Interaction

Scripts

Google mentors interacted with their assigned firms in treatment and control following this provided sample script shown in Figure C.5. The script distinguishes between which content was delivered to participants in both conditions versus only participants in the treatment condition.

— Insert Figure C.5 About Here. —

Verification of Proper Treatment Administration

We engaged in four separate methods to ensure mentor execution of the instructed experimental manipulations.

First, we conducted one-on-one training with each mentor prior to the start of the hackathon to verify comprehension of the experimental treatment (and control) interaction expected of them. In this training, we also practiced the interaction with the mentor, where we role-played as a participant.

Second, a member of the author team spoke to each and every mentor after the mentors completed their visits to their assigned firms. We directly asked each mentor for every firm they visited every single time, "Did you visit table X and how did you conduct the meeting?" After each round of manipulations—three visits per firm spaced out by two hours—we confirmed that they visited all assigned firms and executed the correct experimental manipulation consistent with whether the firm was part of the treatment condition or the control condition.

Third, in addition to directly confirming with each mentor, we also independently verified the statement of the mentor by following each mentor at least once during the competition and viewing the mentor from a distance, i.e., through the door from outside the room where the firms worked.

Fourth, we employed a research assistant who also spoke to every single mentor after each set of meetings. This research assistant also viewed mentor meetings from a distance to verify compliance.

C.1.5 Post Hoc Statistical Power Analysis

For the primary field study, we conducted post hoc power analysis using G*Power's general criteria (Faul et al. 2007). Based on the sample size and the effect size derived from our model estimates, we obtained the retrospective power statistics (1-β) of iterative coordination's influence on innovation outcomes and process.

We began with the power analysis for innovation outcomes measured by *Value* and *Novelty*. Using the output of Model 4.3-1 in Table 4.3, we checked the power of the treatment effect on *Value* with a 5% level two-tailed t-test, the full sample size (38), and the number of predictors (2). Because the study has already taken place, we were able to estimate the

treatment effect size $f^2(0.35)$ using the R^2 (0.26) from Model 4.3-3 ($f^2 = R^2_{partial}/(1-R^2_{partial})$ (Selya et al. 2012)). The power is still above 80% when experimenting with a 1% level two-tailed t-test. Moreover, the power measurement increases in the case of a one-tailed t-test. Similarly for *Novelty*, we estimated the treatment effect size f^2 (0.13) using the R^2 (0.12) from Model 4.3-7. Repeating the post hoc power analysis with a 10% level one-tailed t-test, the full sample size (38) and the number of predictors (2), we obtained a power estimate of 82.5% for Model 4.3-5. However, when choosing a 5% level two-tailed t-test, the power is only 59%. Thus, provided that the assumptions outlined above are appropriate, the slightly less significant statistic $[t(35) = -1.79]$ obtained in the regression analysis outlined in Model 4.3-5 might in fact be due to Type II error. We detail our power statistic result for the two innovation outcomes in Table C.1.

— Insert Table C.1 About Here. —

We then analyzed the power statistics on innovation process measured by the dependent variables *Code Integration Action* and *Advanced API Specialization*. The balanced firm-minute panel data allowed us to run a repeated measures post hoc analysis as a mixed model. We obtained very large power estimates for both innovation process measurements. In fact, for any reasonable effect size (the commonly used G*Power's effect size is 0.2, 0.5, and 0.8), we were able to obtain a power estimate of nearly 100%, as shown in Figure C.6. The primary reason could be due to the large number of repeated measures present in our data (540). As a matter of fact, McKenzie (2012) underlines that multiple measures over time significantly improve experimental power. The author suggests that, when outcomes are noisy and with low autocorrelation, collecting multiple measurements at relatively short intervals generates big gains in experimental power, albeit at a marginally decreasing rate. As a consequence, since our study includes multiple observations over time, we are confident that our regression analysis of firm process from software code has strong statistical power.

— Insert Figure C.6 About Here. —

C.2 Appendix: Outcomes from Primary Study (Software Field Experiment)

We document of a number of additional tests to verify the robustness of the main results presented on the firm outcomes.

C.2.1 Validating Measures of Value and Novelty

We validate the measures of *Value* and *Novelty* measures used in the primary field study and also in the follow-on laboratory study. We consider content (face) validity, convergent validity, and divergent validity. For convergent validity, we show that our measures overlap with identical or closely similar measures in the extant literature. For divergent validity, we show that existing measures and our measures of value and novelty are distinct constructs from one another. For the purposes of this study, the key need is to show that value and novelty are distinct concepts from one another, since the premise of the study relies on that distinction, although they might overlap with other constructs that are not a part of this study, e.g., beauty, arousal.

Content (Face) Validity

To assess the face validity of our two main measures of *Value* and *Novelty* and of the other measures to be used in a questionnaire to assess convergent and discriminant validity (described later), we sent a broad set of possible measures, including these two, to four social sciences researchers to gauge their initial reaction before executing on a quantitative survey study. They were asked to comment on the content, clarity, and scaling of the instruments.

We then conducted pretest interviews with seven product development experts and software engineers to assess whether all items were understandable and clear. Specifically, all the test participants were asked to choose what construct to which each questionnaire item could be referring among the provided options *Value* and *Novelty*. For all the survey items we constructed, nearly all participants answered as we ex ante anticipated across all measures. In particular, all participants correctly identified the correct construct when

queried about our specific formulations of questions for *Value* and *Novelty* that are used in the two experimental studies, thus validating the use of these specific questions to measure what we intended to measure. For the few discrepancies we observe on the other measures, we sought advice from the participants and resolved any ambiguity in the measure. We made minor changes to alternative measures, but not to what would be our ultimate measures of *Value* and *Novelty*, as a result of this feedback. After several iterations of item editing and refinement, the questionnaire was administered to our full sample of survey respondents described in the following section.

The use of these specific *Value* and *Novelty* criteria to evaluate software applications had also been validated and recommended by our co-sponsor, Google, from their experience hosting dozens of prior hackathons, suggesting that these measures hold external validity in our context of software development hackathons.

Survey Design and Sample

To assess convergent and divergent validity, we conduct a survey of 20 people that asks them to evaluate a series of five mobile applications on several constructs drawn from prior literature that relate to our interpretation of value and novelty. Following the practice suggested by Amabile (1996), each survey respondent assessed the set of mobile applications independently in a different randomized order and looked over all products before rating.

To develop the questionnaire, we carefully examine the organizational innovation and product design literature. We document a variety of existing measures across the literature that measure concepts closely related to measures of *Value* and *Novelty*. Table C.2 details the final set of existing measures that emerged in this review, along with the two we eventually used. These measures are used in the questionnaire. For *Value*, we include a battery of identical or closely related concepts that should be convergent with our measure of *Value*: *Relevance* (Eisenberger and Rhoades 2001, McCarthy et al. 2018), *Meaningfulness* (Im and Workman 2004), *Effectiveness* (Rijsdijk and van den Ende 2011), *Business Value* (Girotra et al. 2010), *Overall Quality* (Besemer 1998, Paletz and Peng 2008, Rietzschel et al. 2006,

2010, McCarthy et al. 2018), and *Usefulness, Display, Appropriateness* (Miron-Spektor and Beenen 2015). For *Novelty*, we include: *Novelty* (Miron-Spektor and Beenen 2015), *Originality* (Shalley et al. 2004, McCarthy et al. 2018), *Radicalness* (Mueller et al. 2012), *Newness* (Im and Workman 2004), and *Uniqueness* (Miron-Spektor and Beenen 2015). The majority of these questions are exactly the same as the survey questions used in the prior work; in some cases, we made slight modifications of the original measures to align the wording of the question with others in the survey to adjust it for our particular context (i.e., mobile software applications). Each participant received the same set of instructions and measurement items.

The *Value* measure included seven items, with each related to one of the concepts listed in the *Variables* column of Table C.2, e.g., Girotra et al. (2010) specify *Business Value* as "the use of the idea to a commercial organization that could develop and sell the products." Similarly, the *Novelty* measure included seven survey items, also listed in the same column of Table C.2, e.g., Shalley et al. (2004), McCarthy et al. (2018) specify *Originality* as "the degree to which the idea is different than other available ideas."

— Insert Table C.2 About Here. —

Using these questions, participants are asked to evaluate mobile application wireframes. A wireframe is a type of visual prototype used by mobile application developers to communicate the visual design and functionality of an application (see Ulrich et al. (2020) for additional detail on the use of visual prototypes in the product development process). Wireframes are used to help product development professionals evaluate the quality of a proposed product. Accordingly, products developed at hackathons and other early-stage innovation settings often use wireframes to help judges evaluate the quality of products produced. Given this context, we believe a sample of mobile application wireframes are an appropriate set of products to help establish construct validity.

We ask participants to evaluate five mobile messaging application wireframes. Given the large number of questions posed, we limit the number of wireframe application to five

to ensure the survey could be reasonably completed in under fifteen minutes. Furthermore, we intentionally focus on mobile messaging applications to facilitate a relative comparison of the applications. Given the widespread use of mobile messaging applications today, survey participants have had sufficient exposure to understand what features and visual designs would be novel or valuable. Finally, to ensure that participants did not conflate their ratings with perceptions of familiar brands or existing applications, we ask them to evaluate a new set of applications that do not currently exist on the market.

We recruited 11 MBA and undergraduate students who received training in the business of mobile application development through a series of courses or student-run seminars. In addition, we also enrolled nine subjects who were representatives of the target user market for the mobile applications that the firms developed.

Convergent and Discriminant Validity

To assess convergent and discriminant validity, we apply Confirmatory Factor Analysis (CFA), a commonly applied statistical procedure for testing a hypothesized factor structure (Bagozzi et al. 1991, Segars and Grover 1993, Byrne 2001). For the purpose of assessing the validity of *Value* and *Novelty* measures, we ran CFA for each multi-item scale on all survey items for *Value* and *Novelty* in a Structural Equation Model (SEM) that includes two latent variables—one for the category of measurement variables related to value and another for the category of measurement variables related to novelty—and allows for covariance between these two latent variables. Based on the generated model fit statistics, the specified model appears to fit sufficiently well ($\chi^2_{(76)} = 96.795$, RMSEA = 0.053, AIC = 4232.290, CFI = 0.977, IFI = 0.972). In particular and most importantly, we find that our measures of *Value* and *Novelty* indeed correlate highly ($\rho > 0.5$) with previously used measures in the literature that seek to measure similar concepts, suggesting that our measures align well with constructs in past work that should be related.

After validating the fitted measurement model, we were able to assess the convergent and discriminant validity of the measures of our interest (Segars and Grover 1993). We find

that the squared correlations (SC) among the two latent categories of variables is 0.017. The average variance extracted (AVE) by the latent variables is 0.624 for the latent value category and 0.629 for the latent novelty category. Based on commonly used standards for interpreting SC and AVE (Fornell and Larcker 1981a,b), these assessments suggest that there is no problem with either convergent validity and discriminant validity in the model we put forth.

C.2.2 Starting Goal Analysis

We conduct a post hoc analysis to explore whether the experimental treatment differentially affects the evolution of the goals over time in terms of their value and novelty. For this analysis, we obtain additional data on the starting goals of the participating firms that was not available to us when conducting the primary analyses reported in the main manuscript. Collecting this data also allows us another way of validating the randomization of the firms assigned to treatment and control conditions: absent our experimental intervention and with sufficient randomization, there should be no statistically significant difference in the mean value and novelty of the starting goals between firms in the treatment condition and those in the control condition.

Data

We obtain and code data on the starting goals of the firms participating in the hackathon. By starting goal, we mean the initial description of a novel and valuable technical product that they wished to create by the end of the competition, near the start of the competition, and prior to any experimental intervention. At the start of the hackathon immediately after registration, Google representatives, who were neither mentors nor judges, went around and spoke to every participant as a part of their effort to ensure all participants had access to the technical resources they needed—in terms of computing equipment, software development tools, and troubleshooting advice—in order to successfully develop an application during the competition. Google undertook this effort on its own volition because it wanted to ensure that all participants would be well-positioned to feel a sense of pride and

accomplishment in completing a software development project within the tight timeline of the hackathon competition.

As a part of this effort, these representatives had to inquire what software application each firm wanted to build in order to get a sense of what the technical resource needs of the firm might be. Using a separate tracking system available to Google but previously unavailable to the researchers, these representatives took notes on what the planned software application idea was and what the needs of those applications were. We went back to Google and inquired about access to these proprietary records, and we were able to successfully negotiate access to this data. Thus, through these notes taken by these Google representatives, we have information on the starting goals of the each of the firms in the hackathon.

Variables

We code these goals using largely the same methods as in the judging process at the end of the hackathon, except that the coders in this case look at a text description, whereas the judges in the competition evaluated actual software applications. Two experienced software developers, with expertise in application development, product management, and the advanced Google toolkit of the competition, reviewed each starting application description and scored the descriptions using the same criteria for *Value* and *Novelty* as described in Table 4.1 of the main paper and on the same five-point Likert scale as used by judges at the end of the hackathon. We average the ratings across the two coders ($\alpha = 0.80$). We refer to these new measures as *Starting Goal Value* ($M = 2.21$, $SD = 0.21$) and *Starting Goal Novelty* ($M = 2.79$, $SD = 0.21$).

We find no statistically significant difference between firms in the treatment condition and control condition in their *Starting Goal Value* ($t = 0.44$, $p = 0.66$) or *Starting Goal Novelty* ($t = -0.19$, $p = 0.85$). The comparability of these measures across the two conditions provides further confidence in the efficacy of the experimental randomization.

Methods

We use *Starting Goal Value* and *Starting Goal Novelty* as moderators in our analysis of firm outcomes. We interact these two measures with the main independent variable *Treatment Group*. We use the respective moderator that corresponds with the dependent variable for the same construct as measured on the final software application as it appeared at the end of the hackathon, i.e., *Starting Goal Value* (*Starting Goal Novelty*) appears as a moderator in regressions with *Value* (*Novelty*) as the dependent variable. We run a series of regressions that mirror the models in Table 4.3 of the main manuscript and vary the inclusion of control variables and sample.

Results

Table C.3 presents the result of this post hoc analysis. As a baseline, across all regressions, we find positive associations between *Starting Goal Value* and *Value* and between *Starting Goal Novelty* and *Novelty* ($p < 0.05$ across all models). This suggests that for control firms, the novelty and value of the starting goal may have some latent effect on the realized performance at the end of the competition, as consistent with prior findings (cf. Berg 2014).

— Insert Table C.3 About Here. —

We now turn our attention to the key coefficients of interests on the interaction terms. We find that *Treatment Group*×*Starting Goal Value* has a null or weakly negative effect (e.g., for Model 2, $b = -0.283$, $p < 0.1$). This suggests that having a high *Starting Goal Value* negatively moderates the effect of the intervention, although the total effect of the intervention (*Treatment Group* + *Treatment Group*×*Starting Goal Value*) is still positive. We interpret this as meaning that treatment firms that have low-value starting goals experience a stronger effect from the intervention, shifting more towards value in their final performance as compared to the value of their starting goal, than treatment firms with high-value starting goals.

As a corollary to the above analysis, we find that *Treatment Group*×*Starting Goal*

Novelty is negative and statistically significant (e.g., for Model 4, $b = -0.467$, $p < 0.05$). This suggests that having a high *Starting Goal Novelty* exacerbates the negative effect of the intervention. In other words, treatment firms that have high-novelty starting goals experience a stronger effect from the intervention, shifting more away from novelty in their final performance as compared to the novelty of their starting goal, than treatment firms with low-novelty starting goals.

In summary, this post hoc analysis provides suggestive evidence that there is a boundary condition to the effect of the iterative coordination experimental intervention. Firms that start with a goal that is already high in value and/or low in novelty are not as impacted by the intervention, consistent with a simple intuition that those firms are bounded in how much more value and how much less novelty their final output can be from their starting goal. In contrast, we do not find evidence that iterative coordination reinforces or augments a starting goal that is highly novel, suggesting that the iterative coordination practice generally pushes firms away from a starting goal of novelty and towards value as realized in their final output.

C.2.3 Nonlinear Estimation: Ordered Logit

For robustness, we conduct an additional analysis of firm outcomes utilizing an ordered logit model. Our measures of *Value* and *Novelty*, based on underlying Likert scores on a scale of 1 to 5, are ordinal rather than continuous. Thus, it may be appropriate to utilize a nonlinear estimation method such as ordered logit, which accounts for dependent variables such that the relative ordering of response values is known but the "distance" between them is not. For parsimony, we chose to report OLS estimates in the main paper in Table 4.3. To verify the robustness of those results, in Table C.4, we utilize the same dependent and independent variables used in the models in Table 4.3, except that we estimate the models using ordered logit regressions. We confirm the direction and statistical significance of the reported coefficients using this alternative model.

— Insert Table C.4 About Here. —

C.2.4 Differences in Productivity: Project Completion

We now test whether our results may be explained due to differences in productivity caused by our experimental treatment. We measure productivity in terms of software code generated and completion of their overall software application. Differences in productivity may potentially be explained due to the perceptions of authority. We address this alternate mechanism below.

A potential perception among our hackathon participants of mentors as authority figures is neither likely nor theoretically necessary to enable our iterative coordination treatment. Advantages of traditional authority-based hierarchy include efficiency in coordinating tasks, yielding increased productivity (Lee and Edmondson 2017). Nonetheless, Lawrence and Lorsch (1967) argue that non-hierarchical coordination may be vested in a designated coordinator or integrator who has no formal authority over the individuals whose activities require coordination. Following this approach, we designed the mentor role to serve as facilitator rather than authoritative supervisor of the iterative coordination treatment.

If perceived authority were to be vested in mentors, we would anticipate differences in net productivity among treatment and control firms as the primary observable effect of perceived authority. However, with authority-less mentors, we would anticipate no observable differences in net productivity between treatment and control firms. We test for a possible effect on productivity in outcomes here. Later in Appendix C.3, we test for an effect in the processes, where we evaluate productivity in the writing of software code.

We verify that our experimental treatment did not have a statistically significant effect on whether or not firms completed their software application project by the end of the experiment. In Table C.5, we utilize the same regression models as in Table 4.3 of the main paper, except that we use the dependent variable of *Completion*. The same judges assessed *Completion* on a five-point Likert scale in response to the question: "How far was the firm able to get towards completing and implementing the project?" There is no statistically significant relationship between *Completion* and our experimental treatment,

and the standard errors bound the point estimates within zero.

— Insert Table C.5 About Here. —

Visual Representation

In Figure C.7, we visually present the coefficient estimates of *Customer Needs* and *Novelty* from Table 4.3 of the main paper and *Completion* from Table C.5 of the Appendix; we display coefficient estimates from the odd-numbered models of these tables.

— Insert Figure C.7 About Here. —

C.2.5 Selection into Evaluation

We now verify that our experimental treatment did not have a statistically significant effect on whether firms underwent evaluation, which accordingly suggests that any effect of our experimental treatment on whether or not a judge evaluated a firm did not drive our main finding on firm outcomes in Table 4.3. Consistent with the standard procedure of most hackathons, participants may choose to opt out of participating in the final assessment by the judging panel, and thereby remove themselves for consideration from the set of available awards. A potential explanation for the choice to opt out is the perception by the participating firm that they are unlikely to win any of the available awards, given the firm's *ex ante* private information about the quality of their project. To provide additional detail on the nature of the choice to opt out, we provide additional descriptive analysis in Table C.6 to document whether the choice to opt out may relate to our experimental treatment or observable characteristics of the participants. We find no statistically significant relationship between *Evaluation* and our experimental treatments, nor with any observable characteristics of the firms.

— Insert Table C.6 About Here. —

C.2.6 Moderating Firm Characteristics

To identify boundary conditions and verify the robustness of our firm process analysis, we explore potential moderator variables. We focus on the potential moderators of *Firm Size* and *Graduate Degree*, part of the time-invariant firm characteristics used primarily as control variables in the firm outcomes analysis. These two variables fit particularly well with aspects of our theory and setting. We consider how they affect the role of iterative coordination in both firm outcomes here and firm process later in Appendix C.3.

For *Firm Size*, prior work establishes that coordination costs increase with organization size and, in turn, that coordination costs limit integration activity by the organization (Van de Ven et al. 1976, Okhuysen and Bechky 2009). Thus, we would expect that firms with larger *Firm Size*, as compared to smaller, should benefit more from iterative coordination in terms of undertaking more integration, i.e., *System-Level Branching* and *Code Integration Action*. Based on our theory, this integration should translate to *Value*. On the other hand, a reduction in coordination costs may also translate to performance in novelty because specialized knowledge must be eventually integrated into the final product.

We consider *Graduate Degree* to represent the degree of specialized knowledge present in the firms: higher educational institutions select for those with knowledge and improve the knowledge of those who receive it. Our main process result suggests that iterative coordination is negatively associated with specialization. We expect that firms with greater *Graduate Degree* because they have more specialization "to lose," should engage in even less specialization, e.g., *Subsystem-Level Branching* and *Advanced API Specialization*.

We present results of this analysis in Table C.7. Given the limited sample size, statistical significance is difficult, but there are several key results to highlight. Verifying the robustness of our main findings, we preserve the significant positive and negative effects of the iterative coordination treatment on *Value* and *Novelty*, respectively, in the base terms (Columns 1–4), while we continue to not find a statistically significant effect on *Completion* and *Evaluation* (Columns 5–8). Second, in the full *Novelty* specification in Column 4, we

find that *Treatment Group* × *Firm Size* has a positive effect on *Novelty* ($p < 0.05$), although this same estimate is not significant in the more parsimonious Column 3 specification but the estimate stays the same direction. Nevertheless, the interpretation of this result would be that firms with iterative coordination improve in novelty as their size increases, but the opposite is true for firms without. Naturally, an increase in firm size mechanically introduces specialist knowledge to the firm, but it also introduces coordination costs that can limit whether that specialist knowledge translates into novelty in the final output. Third, we find that *Firm Size* is negatively associated with *Completion* for firms without iterative coordination, but positively associated with those that do. This result further suggests that the role of alleviating coordination costs may be important and, in turn, iterative coordination could be more valuable for organizations with high coordination costs, i.e., larger ones. We do not find any statistically significant interactions for the dependent variable of *Value* (Columns 1–2), nor for the *Treatment Group* × *Graduate Degree*.

— Insert Table C.7 About Here. —

C.3 Appendix: Processes from Primary Study (Software Field Experiment)

We document additional tests to verify the robustness of the main results presented on the firm process analysis. We utilize the same dependent variables as described in the main manuscript in Table 4.4 along with two additional measures described below.

C.3.1 Software Code Hierarchy

We now describe two additional measures of integration and specialization beyond to further confirm the robustness of the results reported in the main paper, which reported on the dependent variables of *Code Integration Action* and *Advanced API Specialization*, respectively. These two separate empirical measures derive from how file hierarchies in the software code reveal underlying knowledge creation in software development.

We take a knowledge-partitioning perspective, which facilitates the interpretation of

knowledge integration and specialization in problems of coordinated exploration (Knudsen and Srikanth 2014). In software development, or product and strategy development more generally, specializing members must coordinate their search efforts and integrate their individual knowledge bases to identify high-performing architectures (Eisenhardt and Tabrizi 1995, Grant 1996). In problems of coordinated exploration, a member's knowledge is represented as a set of partitions of the overall knowledge state-space (Samuelson 2004). Individual members create knowledge such that "the more the partitions in a member's information structure, the greater his or her knowledge about the space" (Knudsen and Srikanth 2014, p. 417). Search, then, proceeds by "going through the current information partitions or if necessary by further partitioning the information structure." Taking this knowledge-partitioning perspective, we apply it to our context of software development.

In software development, developers instantiate knowledge partitions via the creation of partitions in the file hierarchy through directories and files. File hierarchies consist of two types of "nodes": directories, which contain a set of files, and files, which contain a set of code lines. Directories are combinations of constituent files, and each file represents a unique combination of lines of code, embodying a unit of knowledge. File hierarchies, such as the example in Figure C.8, map to design hierarchies as described in Baldwin and Clark (2000). Developers categorize lines of code into files and groups of files into directories to engage in the practice of "information hiding," where elements of a computer program that are most likely to change are purposefully segregated from the rest of the software program (Parnas 1972, Baldwin and Clark 2000). Thus, through information hiding, hierarchies segregate "visible information" from "hidden information," signaling "who has to know what" (Baldwin and Clark 2000, p. 75). In summary, the property of information hiding allows us to map the file hierarchy to abstract knowledge partitions.

— Insert Figure C.8 About Here. —

We distinguish between integration and specialization by documenting evidence for knowledge creation at the system level and the subsystem level of the file hierarchies, respec-

tively. We argue that the level of knowledge creation at different levels of the file hierarchy signals the intended function of a module of code, exploiting a known benefit of modularity in general (Ulrich 1995).

To observe integration in the file hierarchy, we construct *System-Level Branching*, measured as the branching factor specifically for the top-level directory.[69] "System-level" files and directories are closest to the top-level directory, also known as the "root" directory. The average branching factor, a standard performance measure in the computer science literature (Knuth and Moore 1975, Baudet 1978, Muja and Lowe 2009), is the ratio of the number of files and directories below the focal directories at a given time. At the system level, there is a single focal directory, i.e., the top-level "root" directory.

High *System-Level Branching* reflects integration efforts. Files and directories at this level of the hierarchy serve an integrative role, clustering member attention and recombining elements of lower-level files. High-level files and directories are the most visible to all members across a software development firm. Members interface with and look at code starting from the root directory. Thus, for a member to draw attention from other members to her own code, she would place it closer to the top of the file hierarchy. In the broader knowledge state-space represented by the overall file hierarchy, system-level files and directories are a natural place to cluster system-wide attention in line with integration (Okhuysen and Eisenhardt 2002). Furthermore, system-level files and directories recombine elements of lower-level files (Baldwin and Clark 2000), performing a key function of integration (Henderson and Clark 1990, Albert 2018).

To observe specialization in the file hierarchy, we measure *Subsystem-Level Branching*, which calculates the average branching factor for sub-root directories across the file hierarchy, where specialized knowledge creation occurs. "Subsystem-level" files and subdirectories reside further down the file tree, below the root directory. *Subsystem-Level Branching* in-

[69] This measure is functionally equivalent to the number of files and directories that exist immediately below the root directory.

creases if firms increase the number of files per subsystem directory.[70]

Subsystem-level files and directories reflect an intention for information hiding consistent with specialization efforts. Knowledge created at a lower level "only affect[s] their own piece of the system, hence they can be changed without triggering any changes in distant parts of the system" (Baldwin and Clark 2000, p. 75). Thus, to avoid creating or exacerbating existing interdependencies, subsystem-level nodes form a natural place for specialization in knowledge creation. In the absence of coordination, members specialize by developing files at subsystem-level directories. For instance, in a perfectly uncoordinated organization of autonomously searching members (Gavetti et al. 2005), each member would own a personal subdirectory below the root directory within which they would conduct individual search. Figure C.8 illustrates a file hierarchy and the calculation of *System-Level Branching* and *Subsystem-Level Branching*, reflecting integrative versus specialist software development, respectively.

In the context of software development, knowledge integration at the system-level entails key parts of the architecture that touch all parts of the product. In contrast, specialization at the subsystem-level addresses a subsystem or smaller module or component on the product. An example of the relationship between system-level and subsystem-level is Microsoft Windows (system) and the Notepad application (subsystem), which has always been a core part of the overall Windows product. Thus, the terms here refer to the level at which the product is being altered.

Table C.8 provides the summary statistics for these two variables.

— Insert Table C.8 About Here. —

Interpreting "File Hiding"

When agents create files, which they hide, i.e., store, at lower levels of the file hierarchy, we interpret this as evidence of specialization. Lower-level files arise naturally when

[70] *Subsystem-Level Branching* has a minimum value of 1. Holding the number of files (or directories) constant, more directories (fewer files) increases the denominator (decreases the numerator), causing the measure to asymptotically approach a value of 1.

individuals specialize in the development of their own code, without coordination, regardless of whether the code is functional or ultimately used. In contrast, integration takes place at the higher level of the file hierarchy, where those higher system-level files serve as evidence of integration and require coordination among contributors of specialized lower-level files. This choice is a type of information hiding, a broader concept from computer science, whereby engineers hide lower-level specialized components to shift focus on integrating components at a higher level. At least in the context of software development, our key interpretational assumption is that the ultimate act of integration is an independent activity from the creation of specialized files. Thus, evidence of specialization is distinct from evidence of a lack of integration. More explicitly, we do not interpret lower-level file hiding as integration or not, and we do not interpret higher system-level files as specialization or not.

We now describe three cases in which lower-level (hidden) files can exist and constitute specialization. These three cases are comprehensively exhaustive in our setting. First, the vast majority of lower-level files serve a present active purpose, and they reflect specialization in the act of their hiding when they were created. The content of these files may then be integrated by higher-level files to be tied into the main project and used for a present active purpose. Second, some lower-level files contain an error that prevents the current active use. Independent of the error, these error-containing files still reflect the output of specialized activity, although due to time or resource constraints, were not revised to serve a present purpose. Third, some lower-level files may themselves be complete but not yet integrated into the full project. These non-integrated lower-level files may have been intended for integration that was not achieved because of time or resource constraints, or they may be slated for deletion but not yet deleted. These lower-level files were still created through specialization. Even for the files intended for deletion, the intention of deletion does not negate the fact that they were created in the first place.

The relationship between lower-level files in general and integration may be ambiguous. Although we do not use this data in our paper, the presence of the first type of lower-level

file reflects potential integration, while the third type reflects a lack of integration.

Results

We apply the same empirical methodology as used to study *Code Integration Action* and *Advanced API Specialization*, reported in the main document. Table C.9 reports the findings. Model 1 shows that iterative coordination is positively and significantly associated with *System-Level Branching*, such that treatment firms maintained an average of 3.26 more integrative nodes in their file hierarchies in the post-period. To examine the relationship between iterative coordination and knowledge specialization, we turn to Model 2. We find that iterative coordination has a statistically significant, negative association with *Subsystem-Level Branching*, with treatment firms maintaining an average of 0.981 fewer nodes per subdirectory. This indicates decreased knowledge creation at the subsystem level or decreased specialization overall.

— Insert Table C.9 About Here. —

C.3.2 Correlation Table

Table C.10 displays the correlation matrix of these variables. The relatively large correlations of these variables is due to the cumulative nature of the software development process and the long time-series of the underlying data (i.e., at the minute level).

— Insert Table C.10 About Here. —

C.3.3 Standard Errors in Differences-in-Differences Analysis

We present an analysis to verify the robustness of the statistical significance of the main findings reported in Table 4.5 relative to potential inflation of statistical significance. We cluster the standard errors in our main analysis in Table 4.5 at the firm level. While minute-level estimates of our results provide more careful consideration of the dynamics of the search process in software development, the granularity of this data may incidentally inflate the statistical significance of our results. In particular, we may face a serial correlation challenge. Because our estimation relies on a long time-series, our dependent variable likely

serially correlates positively. Moreover, as an intrinsic aspect of a differences-in-differences type model, our key independent variable, *Treatment* × *Post*, by definition changes very little within a firm over time (Bertrand et al. 2004). Therefore, in this supplemental analysis, we collapse the minute-level data to a single observation in the pre-period and a single observation in the post-period, taking the separate averages of the dependent variables for both periods. We confirm the statistical significance of our main findings at the firm-minute level, with identical coefficients and standard errors.

— Insert Table C.11 About Here. —

C.3.4 Effect of Treatment over Time

We also explore whether there are time-varying effects of our treatment. Instead of a single *Treatment* × *Post* indicator variable, we construct three separate independent variables, each representing the interaction between *Treatment* and one of three two-hour-long periods after the initiation of the experimental treatment. Each of these three periods corresponds to the time windows in between the three stand-up meetings and the end of the experiment. We present this analysis in Table C.12.

We confirm the statistical significance ($p < 0.10$) and direction of coefficient estimates for the second and third periods of the experiment for all dependent variables, and we further find statistical significance in the first treatment period for *Subsystem-Level Branching*. These results imply that the observable effect of our treatment is driven largely by differences occurring late into the experiment, suggesting that firms must undergo treatment for a sufficient time or experience a sufficient number of stand-up meetings for observers to recognize a treatment effect.

— Insert Table C.12 About Here. —

C.3.5 Differences in Firm Productivity: Net Software Generated

As anticipated, we find there are no statistically significant differences in the amount of software code written by firms due to our experimental treatment. We measure productivity

in terms of the net software code written by the firms, measured as *Lines*. In Table C.13, we run an analysis resembling our main analysis of firm productivity in Table 4.5, except with dependent variables of *Lines* and $\ln(Lines + 1)$, a logged version of the *Lines* variable to account for skewness in the underlying measure. These results suggest that there are no observable differences in raw software writing productivity due to the experimental treatment.

— Insert Table C.13 About Here. —

We wish to exclude differences in overall productivity that may have occurred due to unobservable heterogeneity. To test the robustness of our main findings on firm process in Table 4.5, we now include the control variable $\ln(Lines + 1)$ to address unobservable time-variant heterogeneity across firms; firm fixed effects already control for unobservable time-invariant heterogeneity across firms. We confirm the statistical significance of our main findings. The statistically significant relationship between $\ln(Lines + 1)$ and the dependent variable is mechanically due to the cumulative nature of the software development process, where a firm in a later period with more *Lines*, relative to the same firm in the earlier period, would have more *System-Level* and *Subsystem-Level Branching*, take on more *Code Integration Action*, and utilize *Advanced API*.

— Insert Table C.14 About Here. —

C.3.6 Meeting Duration and Post-Meeting Latency

For an organization with a fixed or limited amount of time available, the pure act of iterative coordination takes time that the organization could otherwise use for other purposes. In practice, the time that an organization dedicates to its formal coordination meetings comes directly out of the finite time resource of its human capital. An organization that conducts these meetings cannot automatically compel its employees to work longer to make up for the lost time in the experiment. Studying this category of practices requires that we allow

iterative coordination to account for real time. Thus, integrating the intervention time aspect into the experiment would allow us to measure the effect of the experiment on organizational performance in the most realistic way possible.

To provide a sense of what magnitude of time is accounted for by iterative coordination, we provide an estimate of the time effect for each separate meeting and the total. We turn back to the raw software code data and event records. We take the difference between the meeting start times and the first time that each firm enters new software code immediately after the meeting. This difference, *Meeting Duration & Post-Meeting Latency*, represents the total effect of iterative coordination on the time available to each firm to work; each iterative coordination meeting takes up time in both the form of the duration of the actual meeting and the time it takes the firm to regroup and get back to work after the meeting is over.

In Table C.15, we find that there are no statistically significant differences in the *Meeting Duration & Post-Meeting Latency* across treatment and control firms. Nevertheless, the point estimates can provide some sense of magnitude to frame the possible effect. For meetings 1, 2, and 3, there were differences of 5.3, -4.6, and 0.8 minutes, respectively, between treatment and control. The difference in total time taken up by iterative coordination, as opposed to the counterfactual intervention, is about 1.5 minutes: out of the total competition time of nine hours (540 minutes), iterative coordination accounts for less than 0.3% of the total time available to each firm, which we interpret as being relatively small. Thus, in the context of this particular field experiment, any possible time effect from iterative coordination is likely not large or meaningful.

In the follow-on laboratory experiment, we document that the effect of the interruption itself, faced by control firms in this field experiment but not the Condition 1 teams in the laboratory, is also likely not large; see Table 4.8 of the main manuscript for more detail.

— Insert Table C.15 About Here. —

C.3.7 Moderating Firm Characteristics

We consider the effect of these moderators on our analysis of firm process. Table C.16 presents the results of this moderation analysis, where we use *Firm Size* and *Graduate Degree* in an interaction term with the main independent variable. In Columns 7 and 9, we find that *Firm Size* positively moderates the effect of iterative coordination on *Code Integration Action* ($p < 0.01$). In Columns 11 and 12, we find that *Graduate Degree* negatively moderates the effect of iterative coordination on *Advanced API Specialization* ($p < 0.10$). We do not find statistically significant results on the moderator term for the dependent variables of *System-Level Branching* and *Subsystem-Level Branching*, although we do preserve the significance and sign of the iterative coordination treatment.

— Insert Table C.16 About Here. —

C.3.8 Mediation Analysis

We conduct a mediation analysis to empirically measure integration and specialization as mediators between iterative coordination and the outcomes of value and novelty. We follow the practices for measuring mediation relationships as described in the strategy and macro-organizational literature, in contrast to methods applied by micro-organizational and social psychology scholars. Following the guidance of Shaver (2005), we run a generalized structural equation modeling (SEM) analysis using the function as defined in Stata. Our model structure and estimation assumptions are similar to recent work by Kaplan and Vakili (2015), who study the effect of patent characteristics on innovation outcomes.

We combine the firm-time panel data on firm process (capturing processes of integration and specialization) with the cross-sectional data on firm outcomes (capturing the outcomes of value and novelty in the final output of each firm). We combine these into one cross-sectional dataset. We preserve the cross-sectional data as is. For the firm-time panel data, we take the measures of *Code Integration Action* and *Advanced API Specialization* at their final value in the last period of the experiment to capture the cumulative integration

and specialization respectively undertook by each firm over the course of the experiment and reflected in their final output; this measurement optimizes on the relationship between the mediators and the actual output that is assessed by judges to create the outcome measures.

Our SEM model uses the iterative coordination *Treatment* as the main independent variable, *Code Integration Action* and *Advanced API Specialization* as mediator variables reflecting constructs of integration and specialization, respectively, and *Value* and *Novelty* as the dependent variables. We allow for both *Code Integration Action* and *Advanced API Specialization* to be mediators to both *Value* and *Novelty*. While the primary mediation relationship of interest would be of integration to value and specialization to novelty, our model set up allows for integration to be a mediator for novelty and specialization to be a mediator to value, in the interest of completely evaluating all possible relationships. Across all individual regressions in the model, we include the full set of firm control variables used in Table 4.4 of the main paper. We use robust standard errors.

We present the results of the mediation analysis in Table C.17, estimated as one SEM model. The first two columns capture the relationship between the mediators of *Code Integration Action* and *Advanced API Specialization* with the iterative coordination *Treatment*. The last two columns show the effect of the mediators and *Treatment* on the main dependent variables of *Value* and *Novelty*, allowing for mediation in the structure of the first two columns. Consistent with the main paper, we find that *Treatment* has positive relationship with *Code Integration Action* and a negative relationship with *Advanced API Specialization*. Then allowing for this mediation, we find that *Code Integration Action* has a positive and statistically significant effect on *Value*, while *Advanced API Specialization* has a positive and statistically significant effect on *Novelty*. At the same time, *Treatment* no longer has a statistically significant effect on *Value* or *Novelty*, having now been mediated, although we preserve the direction of the signs as in the analysis from Table 4.4 of the main paper.

— Insert Table C.17 About Here. —

In aggregate, these results suggest that there is reason to believe that integration and

specialization are mediators that link iterative coordination to value and novelty, respectively. That said, we recommend caution on interpreting these findings. We do not exogenously vary *Code Integration Action* and *Advanced API Specialization* in our field experiment, and using them as mediators necessarily introduces endogeneity into the analysis. In particular, the strength of the field experiment is our ability to exogenously vary the iterative coordination treatment, implying some degree of causality to our interpretation. We cannot say with the same certainty that these estimates represent a causal mediation relationship.

C.4 Appendix: Follow-On Study (Product Development Laboratory Experiment)

C.4.1 Detailed Experimental Procedure

We recruited 210 participants for a study at the behavioral research laboratory at a northeastern university. They participated in a ninety-minute session: fifteen minutes to obtain consent, complete a pre-experiment survey, issue instructions, and assign teams and locations; sixty minutes for the actual experiment; and fifteen minutes at the end to complete a post-experimental survey, present their products, and vote for a prize winner.

A Priori Statistical Power Analysis

For the follow-on product development laboratory experiment, we conducted a priori power analysis using G*Power's general criteria to calculate the sample size required for our study (Faul et al. 2007). We outlined our result in Table C.18 with the value of our chosen significance level (α) and the power (1-β) estimate.

Notably, from the means and standard deviations of teams in each of the three experimental conditions, we were able to estimate the appropriate effect sizes of the two mechanisns of iterative coordination (opportunities to reprioritize goals and additional interim deadlines) on innovation outcomes and process (*Value*, *Novelty*, *Time to Integrate*, and *Individual Sketches*). Specifically, the cross-sectional analysis statistics in columns 2–4 of Table 4.8 contain all the necessary information needed. We used Cohen's d formula to compute the appropriate effect sizes for each mean-comparison analysis of teams in Conditions 1 vs.

2 and Conditions 2 vs. 3. We calculated Cohen's d as the mean difference between two groups divided by the pooled standard deviation (Cohen's $d = (M_2 - M_1)/SD_{pooled}$, where $SD_{pooled} = \sqrt{(SD_1^2 + SD_2^2)/2}$ (Faul et al. 2007)). After obtaining the effect sizes for all four variables in each of the mean-comparison condition, we proceeded with our a priori power analysis.

We first estimated the sample size required to detect the anticipated difference in *Value* between teams in Conditions 1 and 2 by using a 10% level one-tailed t-test with 80% power, assuming an equal allocation. (The actual allocation is 23 teams in Condition 1, 24 in Condition 2, and 23 in Condition 3.) Here we decided to use a one-tailed t-test because it is reasonable to assume the direction of our treatment effects, given the primary field study outcomes. With the estimated size of effect being 0.84 in this particular setting, we obtained a sample size requirement of 28. Repeating the exercise for *Value* under another setting for teams in Conditions 2 and 3, we obtained a sample size requirement of 50. We continued and changed the pre-specified α and power statistics for the mean-difference t-test for each of the variables of interest. A comprehensive view of the sample sizes required in each of mean-comparison studies is depicted in Table C.18.

— Insert Table C.18 About Here. —

The result of the a priori analysis with the corresponding effect sizes reveals that a sample size of at least 72 is required to achieve a power of 80% and a significance level of 10%. However, as mentioned in Section C.1, the post hoc power analysis for the *Novelty* measure in the field experiment, we obtained relatively low power estimates with a less statistically significant coefficient estimate [$t(35) = -1.79$]. In fact, we found similar results for the effect of the additional question, that is, the comparison between Conditions 1 and 2 in the follow-on laboratory study. As shown in column 5 of Table 4.8, an estimated effect of -0.342 on *Novelty* was observed with a standard error of 0.196 ($p < 0.1$). If we take into account the aforementioned higher chance of committing a Type II error and consequently use a lower power statistic, we would obtain a smaller sample size estimate. For example, if

we conduct a priori analysis for a one-tailed 10% t-test with 65% power, a sample size of only 44 is required. We therefore conclude that the sample of 210 individuals randomly assigned into 70 teams of three individuals in our laboratory experiment would suffice to detect an effect.

Recruitment

We recruited 210 participants, drawn from the general population, to participate in this study in a behavioral research laboratory at a northeastern university. We worked with the Behavioral Research Services (BRS) team in the business school of the aforesaid university to recruit the participants. BRS maintains an ongoing subject pool drawn from a variety of sources: from the students and staff of the university itself; from other local universities or colleges in the same metropolitan area; through social media direct-response advertising; and through traditional online, print (e.g., local newspaper), and in-person (e.g., subway signs) advertising. Interested participants sign up and consent through the Sona Systems research participant management software platform, used by 1,000 universities globally. After signing up, BRS manually assesses participants for consideration in the subject pool and excludes them if they do not meet the following criteria: over 18 years old, fluent in English, and able to accept payment in the United States.

BRS followed the IRB-approved procedures as specified for this particular study. In addition to the BRS qualification criteria, we imposed an additional requirement for our study that all participants must have an education level of at least some college education and a high-school diploma or equivalent. We added this requirement to better target participants who would be able to understand the product development task—design a new dorm or apartment product concept for a manufacturer—and have some sense of how to work in a team.

BRS advertised our study to participants qualified for this study by advertising the session as a post in Sona and in emails to qualifying participants. Figure C.9 documents the posting in Sona and the email template used for individual recruitment.

— Insert Figure C.9 About Here. —

Phase 1: Pre-Experiment (15 Minutes)

Participants enter the lab and are asked to indicate their consent for the study. Then, participants complete a pre-experiment survey on their basic demographic characteristics and their relevant education and experience for the experimental task. After completing the survey, participants are orally informed that they will be randomly assigned to teams of three to complete a product development task; the recruitment materials and aforementioned consent form already inform participants that they will be assigned to teams of three. Participants then receive instructions on their team task, adapted from Girotra et al. (2010):

> *Imagine that you have been retained by a manufacturer of dorm and apartment products to identify new product concepts for the student market. The manufacturer is interested in any product that might be sold to students in a home-products retailer. The manufacturer is particularly interested in products likely to be appealing to students. These products might be solutions to unmet customer needs, or products that have great potential but do not yet exist in the market. These products may also offer better solutions than products that already exist in the market.*
>
> *Your task as a team is to develop one product to present to the manufacturer by the end of today's session. Each of you is provided material to individually sketch and develop your own ideas, which we highly encourage you to use. Note that the most successful products proposals involve input from all team members. We encourage you to consider many product ideas before settling on a final team product. You will be asked to present your final product to your team mentor at the end of today's session using the presentation slide. Only one presentation slide is provided to each team. In addition to sketching and explaining your idea on the presentation slide, be prepared to offer a three-sentence summary of your product idea to your team mentor. Your mentor will then use this three-sentence summary to pitch your product to the rest of the session.*

Phase 2: Experiment (60 Minutes)

After reviewing the task, participants in their randomly assigned teams situate themselves at an assigned team work station, i.e., a private conference room. Each participant receives paper on which to sketch their individual ideas, along with one whiteboard per

team, on which they are required to sketch and describe their final product idea with a three-sentence summary.

Members of the research team act as team mentors who visit the teams intermittently to administer "stand-up" meetings. These team mentors introduce themselves as members of the research team to avoid deception, which is against laboratory policy.

The experiment takes place over several separate sessions, where multiple teams participate in each session. Each experiment session features one and only one of three experimental conditions. Teams within the same session all fall into the same experimental condition and receive the same intervention. The three experimental conditions and their procedures are described in the main manuscript.

At the end of the 60 minutes for the product development task, teams turn over all their individual sketches and their whiteboard with the final product. In addition, teams provide a three-sentence summary of their product design idea, which they are asked to record in bullet points on the whiteboard.

Phase 3 Post-Experiment (15 Minutes)

Participants individually complete a post-experiment survey. After completing the survey, a member of the research team presents each team's product based on the whiteboard sketch and three-sentence product summary collected at the end of Phase 2. Teams in the session then vote on their favorite product; they are barred from voting for their own product. All teams within a session are assigned to the same experimental condition, and so no team is unfairly advantaged through this process. The winning team receives a prize of $10 per person; this prize is in addition to the $25 remuneration for participating in the experiment.

C.4.2 Participant Characteristics and Randomization Check

In a pre-experiment survey, we measure several participant characteristics to confirm the validity of our experimental randomization. Table C.19 presents the summary statistics and an Analysis of Variance (ANOVA) test showing that there are no statistically significant differences in these characteristics.

— Insert Table C.19 About Here. —

C.4.3 Additional Measures

Table C.20 summarizes additional measures collected in the follow-on laboratory study not reported in the main paper. Table C.21 reports the summary statistics and cross-sectional analysis of the measures defined in Table C.20. The following descriptions of results refer to these two tables.

— Insert Table C.20 About Here. —

— Insert Table C.21 About Here. —

Completion

Based on the final project of each team, we code a measure of *Completeness*—following the same coding method for *Value* and *Novelty* described in the main document—with an inter-rater agreement of 0.89. We find no statistically significant difference in the *Completeness* of the final product, and the insignificant point estimate is relatively small.

Coordination and Specialization Post-Experiment Survey

We survey each individual after the experiment. After the experiment, we collect individual retrospective interpretations of the organization of each team and the contributions of each team member. Table C.22 depicts the post-experiment survey questions posted to participants regarding their perception of coordination and specialization within their teams. We adopt a battery of widely used survey measures of coordination (*Effectiveness*, *Few Misunderstandings*, *Low Backtracking*, *Efficiency*, and *Low Confusion*) and specialization (*Group*, *Individual*, *Responsibility*, *Necessity*, and *Awareness*) as proposed by Lewis (2003).We follow exactly the set of questions proposed by Lewis (2003) and used widely in subsequent work. Participants responded on a Likert scale (1–5).

— Insert Table C.22 About Here. —

In Table C.21, we report the results of the self-reported survey-based measures of coordination (i.e., *Coor.∴...*) and specialization (i.e., *Spec.∴...*), we find that a consistent set of estimates suggesting that adding both the question in Condition 2 (versus Condition 1) and the additional meeting in Condition 3 (versus Condition 2) leads to greater self-reported coordination in the teams but lower individual specialization. The direction of the estimates remains consistent throughout these measures, i.e., all positive for coordination and negative for specialization as we go from Condition 1 to Condition 2 to Condition 3. We find statistical significance on several of these survey-based measures, but not on all of them.

Intervention Duration

We use the video recordings to measure the incidental implications of our experimental treatment on the available time of our subjects, as a function of the time spent in the meeting (*Meeting Duration*) and the time after the meeting it takes to get back to work (*Post-Meeting Latency*).

As is mechanically the case, adding an additional meeting leads to more total *Meeting Duration*. However, *Meeting Duration* in Condition 3 of 98.6 seconds (1.6 minutes) is less than double the 68.8 seconds (1.1 minutes) in Condition 2, suggesting there is some diminishing need for formal meeting time as additional meetings are added. The additional question in Condition 2 adds a relatively small amount of time, 29.8 more seconds than in Condition 1 ($p < 0.05$). While these two results suggest there is a time cost to engaging in formal meetings, we must first take this in light of the fact that these differences are relatively small as compared to the total time of the experiment (3600 seconds total), where even in the most saturated Condition 3, it only reflects 3.6% of the total available time ($(98.6 + 31.7)/3600$ seconds).

With respect to the *Post-Meeting Latency*, we do not find a statistically significant difference between Conditions 1 and 2. As expected, the additional meeting in Condition 3 (versus Condition 2) mechanically leads to additional *Post-Meeting Latency* of about 20.9

seconds ($p < 0.05$), but the two meetings of Condition 3 have less than double the *Post-Meeting Latency* time of 32.7 seconds in the one meeting of Condition 2, consistent with the diminishing need pattern we observe on *Meeting Duration*.

Ad Hoc Communication

To understand the effect of our interventions on the ad hoc communication that takes place, we take the video recordings and transcribe the oral communication of each team and identify individual speakers and the time stamps of all communication. We measure the frequency (*Ad Hoc Frequency*) and word count (*Ad Hoc Word Count*) of the oral communication that occurs outside of the interventions.

We find that both the addition of the question in Condition 2 (versus Condition 1) and the additional meeting in Condition 3 (versus Condition 2) leads to a statistically significant increase in the frequency of distinct exchanges, *Ad Hoc Frequency*, that take place outside of the formal meetings ($p < 0.1$ and $p < 0.1$, respectively). Each exchange is an individual's oral communication bounded by speech by other individuals on the team. The estimates on *Ad Hoc Word Count* show no statistically significant differences across the conditions, although the pattern of point estimates goes up as we add an additional question in Condition 2 and an additional meeting in Condition 3.

We find no evidence that iterative coordination reduces the ad hoc communication that takes place outside of the formal meetings. Instead, the result on *Ad Hoc Frequency* suggests that iterative coordination increases the need (or at least realization of the need) to increase communication for interdependent integration purposes, where more meetings essentially reflect greater back-and-forth conversation between the members of the team.

While iterative coordination meetings may theoretically substitute for ad hoc meetings that would otherwise occur in between iterative coordination meetings, our results demonstrate the opposite. That is, instead of finding decreased frequency of ad hoc meetings and decreased volume of ad hoc words exchanged, we find that the frequency of meetings increases for the same (null effect) volume of words exchanged (for Conditions 2 and 3 rela-

tive to that of Condition 1). These results suggest that by updating the goals of teams via more frequent, goal-oriented iterative coordination meetings, the team may democratize its member contributions by involving greater exchanges across the team, for the same volume of words exchanged. In this case, increasing coordination in flat forms may help preserve theorized benefits of autonomy and democratization of contributions (Lee and Edmondson 2017).

Deadline Salience

To uncover details from the oral communication of the participants, we take the raw transcripts of all oral communication generated from the video recordings and apply the widely used text analysis software Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) to identify relevant categories of words used. To generate scores for each category of words, Pennebaker et al. (2015) specify categories of words and then calculates a score for that category reflecting the percentage of words in the transcript of a session that belong to that category out of the total length of the transcript. Recent work by Reypens and Levine (2018) demonstrates the benefit of using text analysis and specifically LIWC as a tool for documenting behavior in strategy research.

As a general measure of the salience of a deadline, we build the measure *Time* to capture words that imply a general focus on matters related to time (e.g., end, until), which we use as a proxy for awareness of the time exhausted and time remaining for the team in a session. An increased attention to time in general is a major consequence of approaching deadlines (Gersick 1988, 1989, Waller et al. 2002).

We then generate two categories of measures related to the two possible theoretical viewpoints we seek to evaluate. The first category of measures intends to capture patterns of communication that would be consistent with the task transition that occurs as a deadline approaches (Gersick 1989, Waller et al. 2002), motivating integration of existing knowledge (Okhuysen and Eisenhardt 2002). *Future Focus* measures the words that suggest a focus on the future (e.g., may, will, soon), which we use as a proxy for attention to deadline.

Second, *Discrepancy* measures the verbs that demonstrate awareness of differences from the status quo, with words such as "would" or "should." Both *Future Focus* and *Discrepancy* words characterize the task transitions that occur as a group approaches the midpoint of time allotted prior to a deadline (Gersick 1988, 1989). For instance, in Gersick (1989)'s laboratory study of groups tasked with creating an advertisement within an hour, a task transition statement such as "maybe we should start on the script pretty soon," which orients the group towards integration of knowledge, features both the *Future Focus* word "soon" and the *Discrepancy* word "should."

The second category of measures intends to capture negative affect reflected in oral communication that might indicate anxiety towards the approaching final deadline. The key measure *Anxiety* measures words that indicate anxiety (e.g., worried, fearful), which in this experiment captures anxiety most likely related to an impending deadline since time is the primary resource constraint on the teams. For robustness, *Negative Emotion* measures a general set of words indicating negative affective processes (e.g., hurt, ugly, nasty), which we use as a general proxy for the general affective state of the individuals on a team. In additional, *Swear* measures the use of informal and vulgar terms, (e.g., damn; other examples intentionally omitted), which we interpret as a more intense expression of negative affect.

We find that both the addition of the question in Condition 2 (versus Condition 1) and the additional meeting in Condition 3 (versus Condition 2) lead to a statistically significant increase in the use of *Discrepancy* words ($p < 0.01$ and $p < 0.10$) and *Time* ($p < 0.05$ and $p < 0.05$) words. This is consistent with our theory regarding how an attention to time and a recognition of discrepancy in current progress would occur with opportunities to reprioritize goals and with additional deadlines. We find null effects on *Future Focus*, which may reflect that words using the future tense do not distinctly characterize deadline saliency. In contrast, we do not find any statistically significant results associated with any condition for *Anxiety*, *Negative Emotion*, and *Swear*.

This pattern of findings is consistent with our theorized mechanisms. In particular,

the additional deadline triggers a shift in activity (measured by *Discrepancy*). Coupled with other existing evidence—faster *Time to Integrate* and the post-experiment survey on coordination and specialization—we interpret this shift as presumably towards integrating knowledge. This shift is triggered by an attention to time, a focus on the future deadline, and a recognition of discrepancy between the current state and the future desired state.

On the other hand, while one might have expected the treatment to increase anxiety among participants, we find no evidence of this. Whereas prior literature in time-constrained innovation settings highlights how anxiety may potentially undermine teams trying to coordinate their work prior to a deadline (Lifshitz-Assaf et al. 2020), we find no such anxiety among our participants. Here, the lack of anxiety may relate to an ability to reprioritize goals using iterative coordination. Prior work demonstrates how unfilled goals create high anxiety in individuals (Masicampo and Baumeister 2011). By reprioritizing goals, individuals treated with iterative coordination may create new plans for unfilled goals, helping reduce their anxiety (Masicampo and Baumeister 2011).

C.4.4 Correlation Table

Table C.23 presents the matrix of pairwise correlations for all measures in the follow-on laboratory study.

— Insert Table C.23 About Here. —

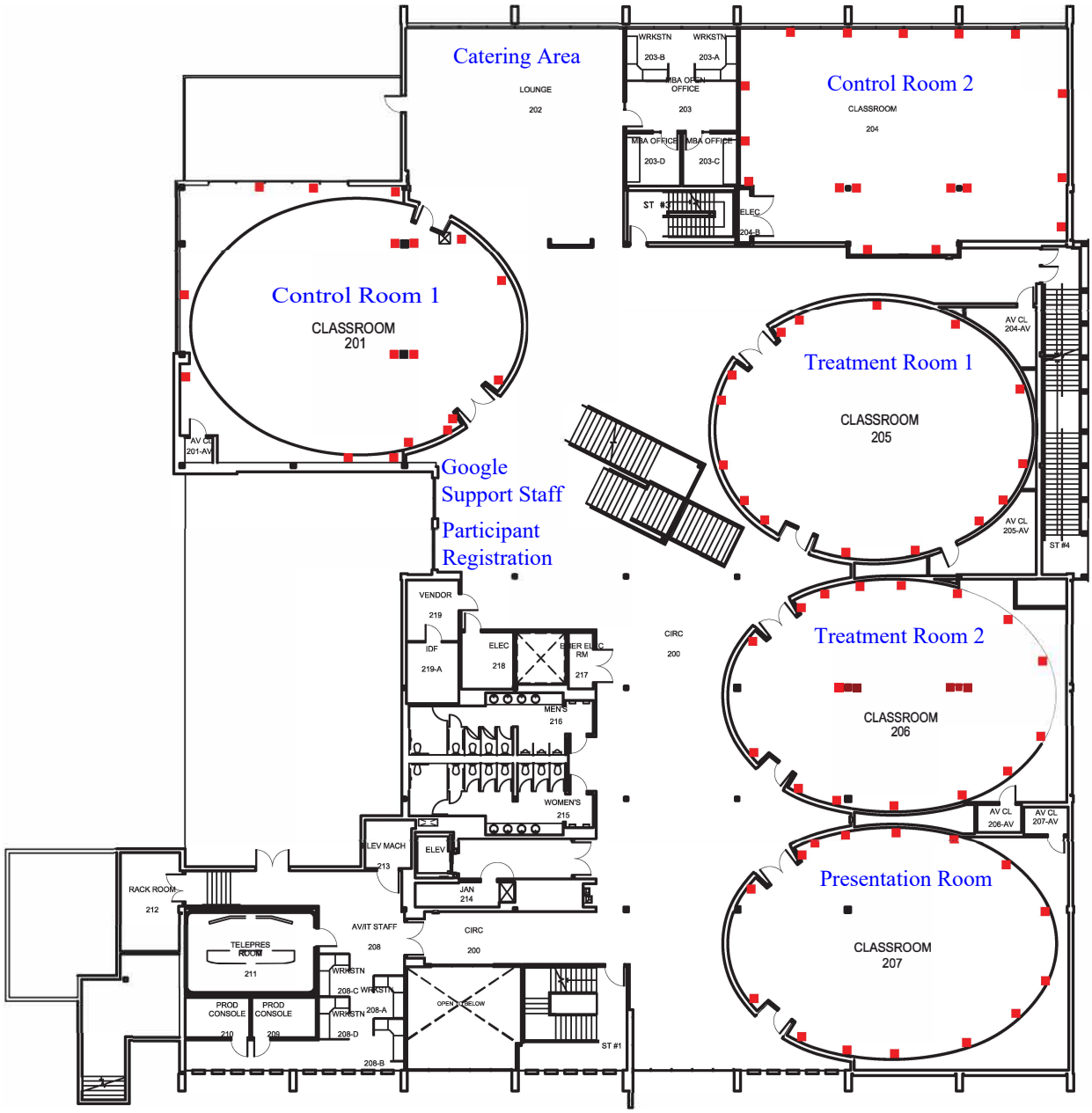
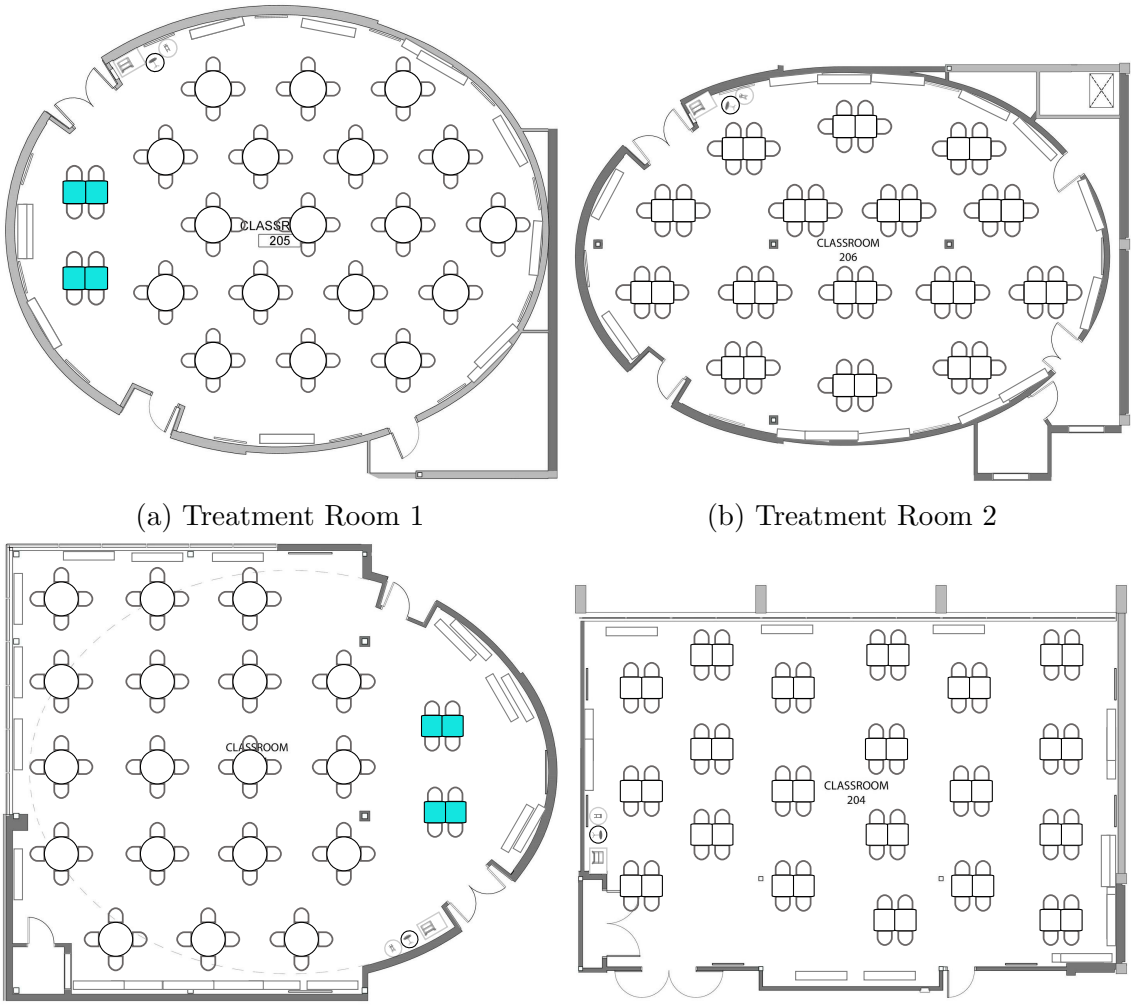


Figure C.1: **Primary Field Study: Overall Floor Plan.** This diagram depicts the overall floor plan of the space where the "hackathon" field experiment took place. Participants registered at tables located at *Participant Registration*. The treatment and control condition where divided into separate rooms, with two rooms for each condition. Lunch and dinner were served in the *Catering Area*; for lunch there was no public seating available, and participants brought their lunches back to eat at the tables where they worked. Staff stored their personal items and rested at the *Google Support Staff* station.



(a) Treatment Room 1

(b) Treatment Room 2

(c) Control Room 1

(d) Control Room 2

Figure C.2: **Primary Field Study: Floor Plan for Treatment and Control Rooms.** These detailed floor plans show the layout of tables where participants work in teams. There are more tables than teams. Participants in the control and treatment condition are assigned to different rooms, so they never observe the mentor/participant interaction of participants in the other condition.

Event Website Description
The team at Google and a researcher from [UNIVERSITY] got together and asked, "How can we work together with the [CITY] tech community to make a difference?"

We're bringing the best in [CITY] tech and research together to write software that makes people's lives better. Come hack on the latest Google Firebase and Cloud, and make an app that does something great.

This event is also a wonderful opportunity for us to understand the software development process. Your participation will help contribute to research on this matter, helping educate start-ups, companies, and other practitioners about ways to make the world a better place with software.

Representatives from Google Cloud and Firebase will be on-hand to offer help in using their tools. We also have guest speakers and other fun stuff planned, so stay tuned :)

Well what are you waiting for? Check out the FAQs below for further details. We look forward to seeing you at the hackathon!

Frequently Asked Questions
What? Hacking for Good with Google.
When? [DATE AND TIME]
Where? [LOCATION]
Prizes? Yes – swag and other cool stuff!
Teams? Teams of 3 to 4 are highly encouraged. If you don't have a team - no sweat! Register below and we'll help you find one :)
How does registration work? We want to bring as many people as possible to this event, but do have space constraints. In order to find a diverse mix of students and professionals, we'll be admitting participants in two waves (see deadlines and notification dates below). [REGISTRATION DEADLINE WINDOWS]
Why attend? Come and program, network and have fun with us!
How can I get more info? Shoot us an email at [EMAIL].

Email Message Template
Dear [NAME],

Would you like to showcase your skills to a leading software development company and make a difference while doing it? Come take part in Google Cloud's first hackathon in [CITY]!

To learn more and register, please follow the link here: [LINK].

If you have any questions, please let me know. Hope to see you there!

Regards,
[SENDER INFORMATION]

Figure C.3

Figure C.3 *(previous page)*: **Primary Field Study: Recruiting Materials.** Solicitation content used to recruit software developers to the hackathon field experiment. For email messages sent to industry professionals on behalf of our co-sponsor, this email message was used as a template. The co-sponsor could customize this message, although they may not promise any additional benefits or features of the competition beyond what is already described in the website.

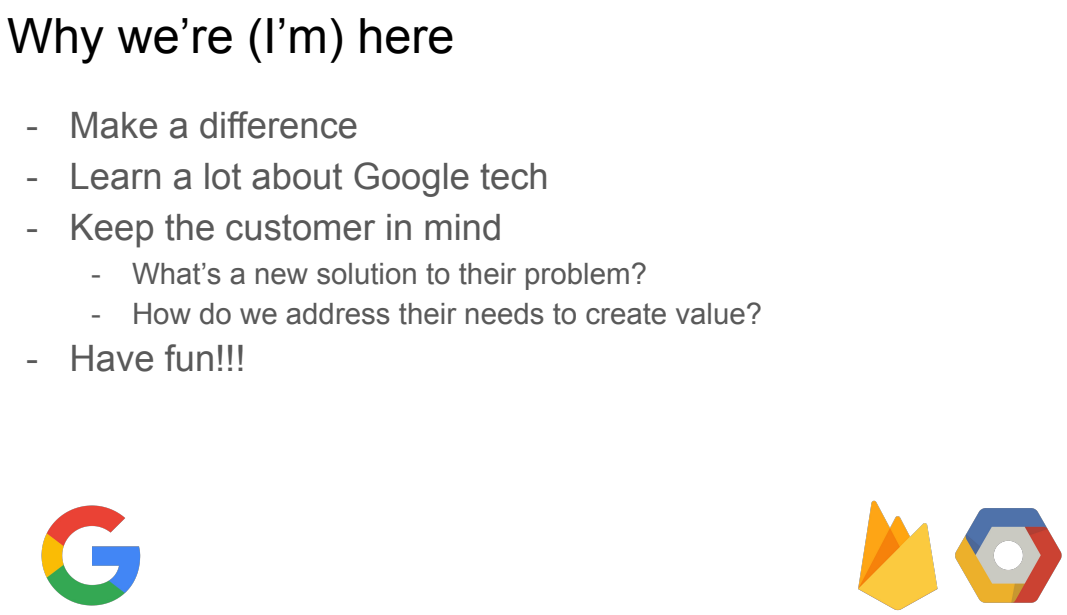


Figure C.4: **Primary Field Study: Competition Guidance.** Slide used in opening presentation by hackathon cosponsor Google to inform participating software developer of the purpose and goals of the competition.

Mentor Introduction (Prior to Competition)
Hello, my name is [NAME] and I'll be your mentor for today! I hope you're excited to participate in today's hackathon. First, let's run through some logistics. To participate in today's competition, we are asking each team to use Github version-control. Has your team successfully set up on Git? [Pause and verify.] Great! We will meet at your assigned table every two hours for a team check-in. Please be sure to be at your table at these times.

[Treatment only.] *At each of these meetings, I will be asking you to consider three check-in questions as a team. These questions are solely meant to help your team's process:*

- *What have you accomplished since your last check-in?*
- *What are your goals until the next check-in?*
- *What are your goals for the end of the day (and have they changed)?*

As mentioned in the welcome presentation, as mentors, we are not involved in the judging process. We are simply here to help with whatever you may need. Let us know if you have any questions! If you can't find me, feel free to ask any Googler for help. Good luck, and see you again at the first meeting!

Meeting 1
Hello there! How's it going? I'm here for the first check-in. Are you enjoying the hackathon so far? [Pause for response.] Excellent!

[Treatment only.] *Let's go through the questions I mentioned earlier:*

- *What have you accomplished since the beginning of today's competition?*
- *What are your goals until the next check-in?*
- *What are your goals for the end of the day (and have they changed)?*

Let a Googler know if you have any questions. See you in two hours!

Meeting 2
Hello there! How's it going? I'm here for the second check-in. How was lunch? [Pause for response.] Good!

[Treatment only.] *Let's go through the check-in questions I mentioned earlier:*

- *What have you accomplished since your last check-in?*
- *What are your goals until the next check-in?*
- *What are your goals for the end of the day (and have they changed)?*

Let a Googler know if you have any questions. See you in two hours!

Meeting 3. Hello there! How's it going? I'm here for the third and final check-in. Are you excited for the end of the hackathon? [Pause for response.]

[Treatment only.] *Let's go through the check-in questions I mentioned earlier:*

- *What have you accomplished since your last check-in?*
- *What are your goals for the end of the day?*

Let a Googler know if you have any last-minute questions. Good luck!

Figure C.5

Figure C.5 *(previous page)*: **Primary Field Study: Google Mentor Scripts.** Google mentors interacted with their assigned teams in treatment and control following this provided sample script. The mentors only spoke the portions of script in *italics* when interacting with teams in the treatment condition and not with those in the control condition. Other instructions for the mentor are shown in brackets.

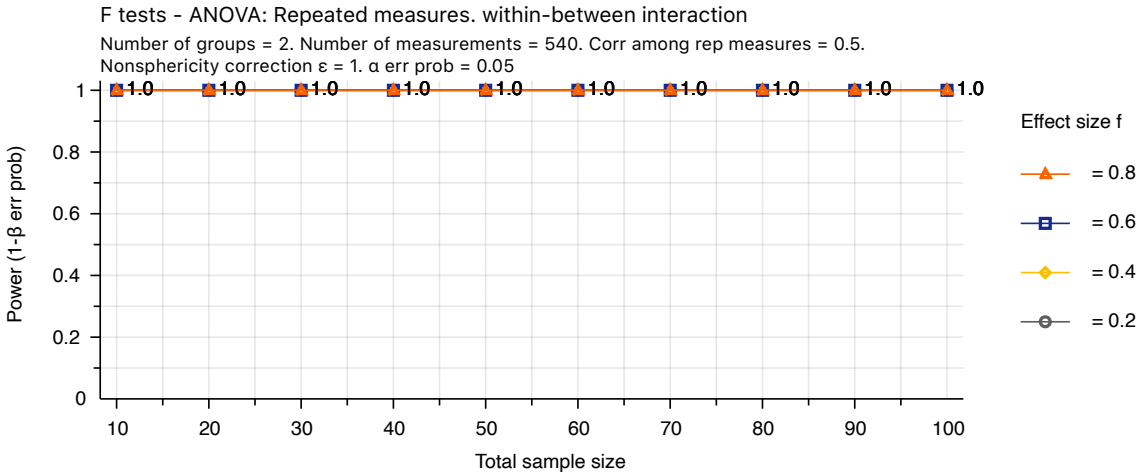


Figure C.6: **Primary Field Study: Power Statistics of Firm Process.** This post hoc analysis was conducted under a 5% level F-test for mixed model, using ANOVA-approach with repeated measures. The figure shows the computed achieved power, given α, sample size and multiple levels of effect size.

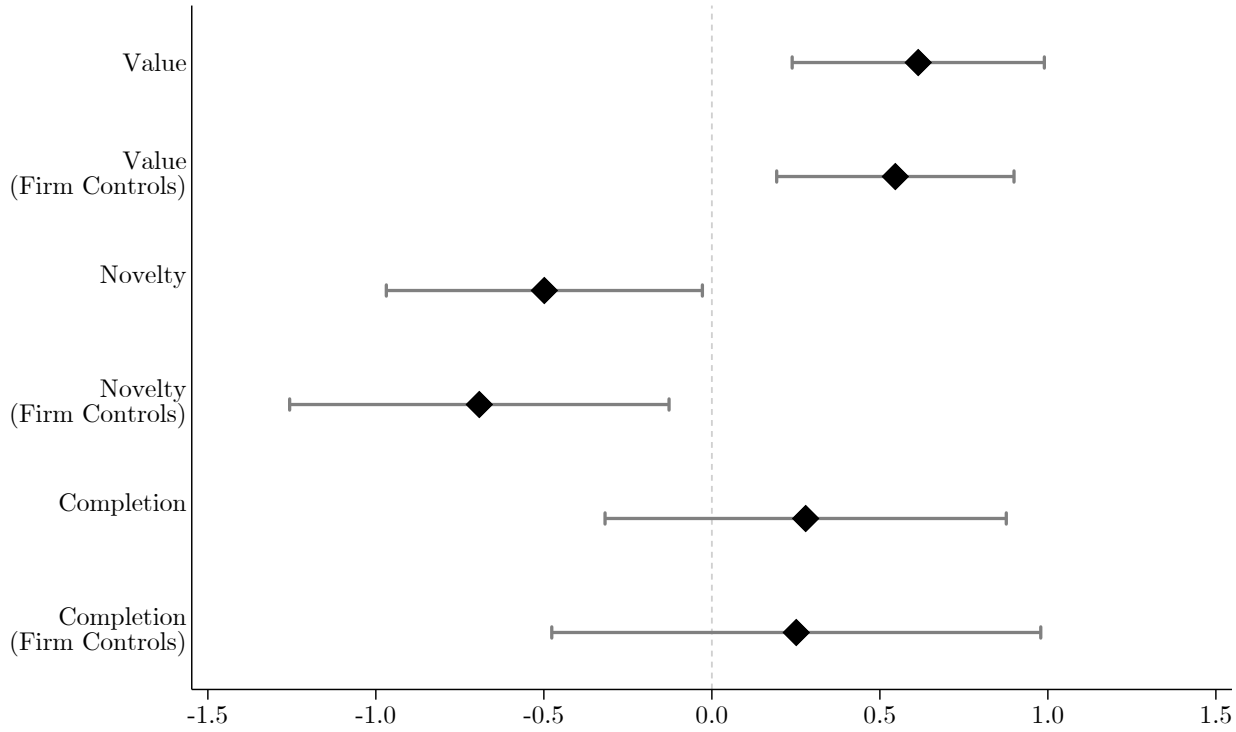


Figure C.7: **Primary Field Study: Firm Outcomes.** This figures plots the point estimates and 90% confidence internals for the analysis of the effect of the experimental treatment on firm outcomes. All estimates include the full set of observations, and the estimates labeled as "Firm Controls" include the full set of firm-level control variables.

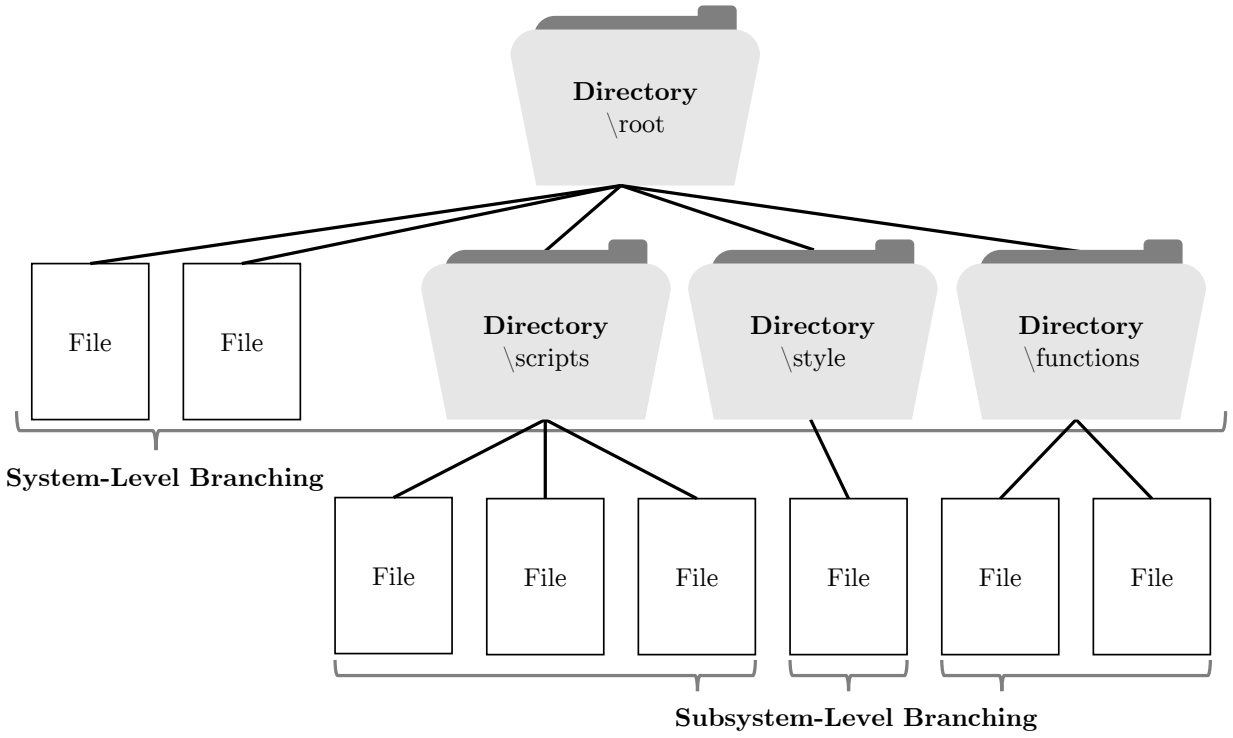


Figure C.8: **System-Level versus Subsystem-Level Branching.** To illustrate the measurement of *System-Level Branching* and *Subsystem-Level Branching*, we present an illustrative file hierarchy. *System-Level Branching* takes a value of $\frac{(2+3)}{1} = 5$, because there are 2 files and 3 directories at the system level, i.e., below the "root" directory, divided by 1 root directory. *Subsystem-Level Branching* takes a value of $\frac{3+1+2}{3} = 2$, because there are $3 + 1 + 2 = 6$ files at the subsystem levels contained within 3 directories.

Sona Session Posting

Study Name: Iterative Coordination Mechanisms

Duration: 90 minutes

Abstract: Laboratory Experiment

Description: The purpose of this study is to build understanding of best practices in product development.

You will be assigned to teams and asked to complete a product development challenge. You will not need any prior skills or knowledge in product development. We don't believe there are any risks from participating in this research. You will be asked to sketch ideas for a new product, in addition to recording your ideas in audio and video. Your data will be kept confidential.

If you agree to take part in this study, we will pay you $25 for your time and effort. We will distribute the payment in cash at the end of the session. In addition, you will have the chance to compete for an additional prize of up to $10!

Preparation: Please bring a government-issued or student photo ID.

Recruitment Email Template
This template was customized for each prospective participant.

Hello [First Name],

We are reaching out to let you know that we have posted sessions for a study (Iterative Coordination Mechanisms) in which you are eligible to participate!

The purpose of this study is to understand best practices in product development. In this study, you will be asked to assigned to teams and asked to complete a product development challenge. You will not need any prior skills or knowledge in product development.

You will be paid $25 for approximately 90 minutes of your time. In addition, you will have a chance to compete for an additional prize of up to $10.

To participate in this study, click on the following link: [Link to Sona Study Page].

Thank you,
[LABORATORY NAME] Administrators

Figure C.9: **Follow-On Laboratory Study: Recruitment Materials.** Posting in Sona and email template used for recruiting participants for experiment.

Table C.1: **Primary Field Study: Power Statistics of Firm Outcomes.** This post hoc analysis was conducted under a single regression coefficient t-test for fixed model, using linear multiple regression. The table result shows the computed achieved power (rounding to the nearest percentage), given α, sample size and effect size.

Variable	α	One-Tailed Test	Two-Tailed Test
Value	10%	99%	97%
	5%	97%	95%
	1%	88%	82%
Novelty	10%	82.5%	71%
	5%	71%	59%
	1%	43%	33%

Table C.2: **Primary Field Study: Related Survey-Based Measures of Value and Novelty in Extant Literature.** These survey measures are extracted from a variety of existing innovation strategy, management, organizational behavior, and product design and development literature. The first questions in each grouping for *Value* and *Novelty* is the same used by judges in Primary Field Study. A survey with a questionnaire using these measures generates the data used to assess convergent and discriminant validity.

Variable Label	Survey Question
Value	*To what extent do you agree with the statements below?*
Value	"The project appeals to the intended market." (This Study)
Relevance	"The project is relevant to the problem at hand." (Eisenberger and Rhoades 2001, McCarthy et al. 2018)
Meaningfulness	"Compared to competitors [existing software applications serving similar purpose(s)], the new product is relevant to customers' needs and expectations." (Im and Workman 2004)
Effectiveness	"The product developed by the project team conforms to performance specifications required by customers and meets the technical requirements of customers." (Rijsdijk and van den Ende 2011)
Business Value	"The use of the ideas to a commercial organization could develop and sell the products." (Girotra et al. 2010)
Overall Quality	"Overall, I deem this product or a selection of its ideas attractive and likable." (Besemer 1998, Paletz and Peng 2008, Rietzschel et al. 2006, 2010, McCarthy et al. 2018)
Usefulness, Display, Appropriateness	"I would like to use the product in my home or office." (Miron-Spektor and Beenen 2015)
Novelty	*To what extent do you agree with the statements below?*
Novelty	"The project helps solve the problem in a new and ambitious way." (This Study)
Novelty	"The product is novel." (Miron-Spektor and Beenen 2015)
Originality	"The idea is different than other available ideas." (Shalley et al. 2004, McCarthy et al. 2018)
	"The idea is original with respect to the unmet need and proposed solution." (Girotra et al. 2010)
Radicalness	"The idea suggests a departure from the current status quo." (Mueller et al. 2012)
Newness	"Compared to competitors [existing software applications serving similar purpose(s)], the new product is really out of ordinary." (Im and Workman 2004)
Uniqueness	"The product is different from other products." (Miron-Spektor and Beenen 2015)

Table C.3: **Primary Field Study: Starting Goal as Moderator.** Ordinary least squares (OLS) estimation of cross-sectional firm-level data. We introduce *Starting Goal Value* and *Starting Goal Novelty* as interaction terms. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Value		Novelty	
Variable	(1)	(2)	(3)	(4)
Treatment Group	1.341***	1.243*	0.883	0.714
	(0.437)	(0.478)	(0.528)	(0.740)
Starting Goal Value	0.398***	0.358*		
	(0.145)	(0.147)		
Treatment Group × Starting Goal Value	-0.276	-0.283†		
	(0.167)	(0.152)		
Starting Goal Novelty			0.540***	0.507***
			(0.119)	(0.140)
Treatment Group × Starting Goal Novelty			-0.464*	-0.467*
			(0.176)	(0.226)
Current Student		0.580		0.181
		(0.346)		(0.440)
Graduate Degree		0.562		-0.020
		(0.347)		(0.468)
Github		0.375		0.839
		(0.626)		(0.872)
Google Development		-0.328		0.715
		(0.401)		(0.488)
Software Development		-0.062*		-0.020
		(0.027)		(0.042)
Prior Hackathons		-0.044		-0.024
		(0.089)		(0.108)
Team Size		-0.174		-0.229
		(0.115)		(0.147)
No Evaluation	-2.953***	-3.199***	-2.460***	-2.524***
	(0.305)	(0.331)	(0.333)	(0.378)
Constant	2.167***	2.503*	1.727***	1.538†
	(0.449)	(0.901)	(0.465)	(0.769)
R^2	0.902	0.939	0.838	0.873
Sample	Full	Full	Full	Full
Observations	38	38	38	38

Table C.4: **Primary Field Study: Regression Analysis of Firm Outcomes Using Ordered Logit.** This analysis resembles the main firm outcome analysis of the paper but instead estimated using an ordered logit model, which has favorable properties for Likert-scale discrete dependent variables. Robust standard errors shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Value				Novelty			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Group	2.310*	2.310*	2.537*	2.537*	-1.383†	-1.383†	-2.989*	-2.989*
	(0.946)	(0.951)	(1.091)	(1.097)	(0.802)	(0.806)	(1.339)	(1.346)
Current Student			4.305†	4.305†			2.370	2.370
			(2.456)	(2.470)			(2.022)	(2.033)
Graduate Degree			1.968	1.968			-2.662	-2.662
			(2.117)	(2.129)			(1.894)	(1.904)
GitHub			1.191	1.191			2.477	2.477
			(4.100)	(4.123)			(2.258)	(2.271)
Google Development			-1.325	-1.325			2.363*	2.363*
			(3.037)	(3.054)			(1.427)	(1.435)
Software Development			-0.159	-0.159			-0.011	-0.011
			(0.149)	(0.150)			(0.102)	(0.103)
Prior Hackathons			-0.738†	-0.738†			-0.019	-0.019
			(0.404)	(0.406)			(0.428)	(0.430)
Team Size			-0.318	-0.319			-1.183*	-1.183*
			(0.616)	(0.619)			(0.563)	(0.566)
No Evaluation	-40.334***		-43.187***		-38.875***		-44.855***	
	(0.860)		(2.513)		(0.684)		(3.050)	
Log Likelihood	-28.29	-28.29	-20.63	-20.63	-35.07	-35.07	-28.06	-28.06
Pseudo R^2	0.488	0.127	0.627	0.364	0.413	0.0484	0.530	0.239
Estimation	Ord. Logit	Ord. Logit	Ord. Logit	Ord. Logit	Ord. Logit	Ord. Logit	Ord. Logit	Ord. Logit
Sample	Full	Evaluation	Full	Full	Full	Evaluation	Full	Full
Observations	38	27	38	27	38	27	38	27

Table C.5: **Primary Field Study: Regression Analysis of Completion.** This analysis assesses the effect of the experimental treatment on the degree of *Completion* of the final submitted software application. Judges assessed *Completion* on a five-point Likert scale in response to the question: "How far was the firm able to get towards completing and implementing the project?" *Completion* has a mean of 2.000 and a standard deviation of 1.660. Ordinary least squares (OLS) estimation of cross-sectional data at the firm level. Robust standard errors shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Completion			
	(1)	(2)	(3)	(4)
Treatment Group	0.279	0.385	0.251	0.336
	(0.353)	(0.484)	(0.428)	(0.594)
Current Student			0.336	0.917
			(0.575)	(0.932)
Graduate Degree			-1.065	-1.640†
			(0.649)	(0.876)
Github			-0.223	0.013
			(1.206)	(1.393)
Google Development			-0.127	-0.352
			(0.617)	(0.922)
Software Development			-0.042	-0.043
			(0.058)	(0.076)
Prior Hackathons			-0.094	0.021
			(0.147)	(0.290)
Team Size			-0.169	-0.195
			(0.208)	(0.300)
No Evaluation	-2.772***		-2.929***	
	(0.257)		(0.306)	
Constant	2.670***	2.615***	4.012***	3.640
	(0.319)	(0.367)	(1.426)	(2.193)
R^2	0.614	0.0249	0.699	0.362
Estimation	OLS	OLS	OLS	OLS
Sample	Full	Evaluation	Full	Evaluation
Observations	38	27	38	27

Table C.6: **Primary Field Study: Regression Analysis of Selection into Evaluation.** This analysis assesses the effect of the experimental treatment on whether firms select into evaluation by judges at the end of the competition. The dependent variable *Evaluation* is an indicator variable taking a value of 1 if the firm underwent judging, and 0 otherwise. The first two models are estimated using ordinary least squares (OLS) regression, and the last two models are estimated using a logit regression. Robust standard errors shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Evaluation			
	(1)	(2)	(3)	(4)
Treatment Group	0.128	0.107	0.539	0.539
	(0.149)	(0.183)	(0.846)	(0.846)
Current Student		-0.293	-1.590	-1.590
		(0.313)	(1.654)	(1.654)
Graduate Degree		-0.050	-0.366	-0.366
		(0.303)	(1.466)	(1.466)
Github		0.111	0.538	0.538
		(0.401)	(1.986)	(1.986)
Google Development		0.288	1.611	1.611
		(0.380)	(1.912)	(1.912)
Software Development		0.003	0.017	0.017
		(0.028)	(0.159)	(0.159)
Prior Hackathons		0.048	0.267	0.267
		(0.101)	(0.535)	(0.535)
Team Size		0.003	-0.056	-0.056
		(0.110)	(0.524)	(0.524)
Constant	0.650***	0.496	0.132	0.132
	(0.110)	(0.638)	(3.001)	(3.001)
R^2	0.0198	0.0923		
Log Likelihood	-23.49	-22.03	-21.01	-21.01
Pseudo R^2			0.0810	0.0810
Estimation	OLS	OLS	Logit	Logit
Observations	38	38	38	38

Table C.7: **Primary Field Study: Regression Analysis of Firm Outcomes Interacted with Firm Characteristics.** Ordinary least squares (OLS) estimation of cross-sectional firm-level data. We introduce firm characteristics of *Firm Size* and *Graduate Degree* as interaction terms; both variables are demeaned to facilitate interpretation of the interaction terms, i.e., the incremental effect of the experimental treatment of *Firm Size* for one additional person, at the mean *Firm Size*. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Value		Novelty		Completion		Evaluation	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Group	0.603*	0.526*	-0.548†	-0.817*	0.261	0.210	0.143	0.157
	(0.259)	(0.227)	(0.276)	(0.330)	(0.350)	(0.484)	(0.167)	(0.190)
Firm Size	-0.069	-0.209	-0.303	-0.772*	-0.456†	-0.606	0.260†	0.290
	(0.305)	(0.208)	(0.225)	(0.320)	(0.236)	(0.375)	(0.148)	(0.202)
Treatment Group	0.102	0.154	0.385	0.878*	0.792*	0.882†	-0.430†	-0.439
× Firm Size	(0.355)	(0.224)	(0.376)	(0.317)	(0.350)	(0.477)	(0.216)	(0.266)
Graduate Degree	-0.166	0.301	-0.515	0.030	-1.345*	-1.133	-0.085	-0.236
	(0.463)	(0.552)	(0.354)	(0.674)	(0.510)	(0.761)	(0.319)	(0.337)
Treatment Group	0.474	0.256	-0.035	-0.775	0.530	0.372	0.182	0.275
× Graduate Degree	(0.718)	(0.603)	(0.683)	(0.843)	(0.883)	(1.022)	(0.443)	(0.471)
Current Student		0.689†		0.114		0.298		-0.228
		(0.363)		(0.371)		(0.601)		(0.331)
Github		0.117		1.330		-0.074		-0.097
		(0.920)		(1.027)		(1.222)		(0.312)
Google Development		-0.152		1.332*		0.159		0.012
		(0.394)		(0.537)		(0.702)		(0.444)
Software Development		-0.050		0.002		-0.039		0.008
		(0.030)		(0.035)		(0.045)		(0.025)
Prior Hackathons		-0.144		-0.061		-0.076		0.033
		(0.089)		(0.127)		(0.134)		(0.093)
No Evaluation	-3.506***	-3.734***	-3.467***	-3.593***	-3.008***	-3.142***		
	(0.245)	(0.209)	(0.223)	(0.270)	(0.315)	(0.378)		
Constant	3.291***	3.440***	3.625***	2.082*	2.838***	3.028*	0.590***	0.699*
	(0.284)	(0.866)	(0.189)	(0.904)	(0.316)	(1.187)	(0.118)	(0.337)
R^2	0.877	0.924	0.793	0.853	0.704	0.725	0.153	0.206
Sample	Full	Full	Full	Full	Full	Full	Full	Full
Observations	38	38	38	38	38	38	38	38

Table C.8: **Primary Field Study: Variable Definitions and Summary Statistics of Firm Process from Source Code File Hierarchies.** Alternate dependent variables constructed from source code file hierarchies over time defined below with their conceptual interpretation. Observations are at the firm-minute level, with 20,520 firm-minute observations across 38 firms.

Variable (Interp.)	Definition	Mean	SD	Min	Max
System-Level Branching (Integration)	Branching factor at the root of the file hierarchy. Equivalent to files and directories directly below the root.	6.271	6.021	1	56
Subsystem-Level Branching (Specialization)	Average branching factor for directories below the root of the file hierarchy.	2.271	1.624	1	7.667

Table C.9: **Primary Field Study: Regression Analysis of Firm Processes from Source Code File Hierarchies.** Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	System-Level Branching	Subsystem-Level Branching
Treatment x Post	3.258*	-0.981*
	(1.413)	(0.426)
Firm FE	Yes	Yes
Time FE	Yes	Yes
R^2	0.410	0.416
Adjusted R^2	0.394	0.400
Firms	38	38
Observations	20,520	20,520
Level	Firm-Minute	Firm-Minute

Table C.10: **Primary Field Study: Firm Process Correlations.** This table shows pairwise correlations for the dependent variables used in the firm-minute level panel analysis of processes.

		(1)	(2)	(3)	(4)
(1)	System-Level Branching	1			
(2)	Subsystem-Level Branching	0.320	1		
(3)	Code Integration Action	0.517	0.385	1	
(4)	Advanced API	0.220	0.302	0.258	1

Table C.11: **Primary Field Study: Regression Analysis of Firm Process at Firm-Post Level.** This analysis of firm process resembles the main analysis of the paper, except that firm-minute observations are aggregated into a single observation in the pre-treatment period and a single observation in the post-treatment period. All dependent variables are averaged over each pre- and post-treatment period of time. Time fixed effects represent indicators for each hour of the experiment. Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	(1) System-Level Branching	(2) Subsystem-Level Branching	(3) Code Integration Action	(4) Advanced API
Treatment x Post	3.258*	-0.981*	2.074*	-1.124**
	(1.413)	(0.426)	(0.878)	(0.408)
Post	3.613***	1.831***	1.532***	1.497***
	(0.984)	(0.358)	(0.420)	(0.361)
Firm FE	Yes	Yes	Yes	Yes
Time FE	No	No	No	No
R^2	0.619	0.552	0.530	0.439
Adjusted R^2	0.609	0.540	0.517	0.424
Firms	38	38	38	38
Observations	76	76	76	76
Level	Firm-Pre/Post	Firm-Pre/Post	Firm-Pre/Post	Firm-Pre/Post

Table C.12: **Primary Field Study: Regression Analysis of Firm Process Allowing for Time Heterogeneity.** This table show the regression analysis using three separate Treatment × Post variables representing each time period after each of the three coordination exchanges that occurred in the treatment period. Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	(1) System-Level Branching	(2) Subsystem-Level Branching	(3) Code Integration Action	(4) Advanced API Specialization
First Treatment x Post	0.749	-0.738†	0.255	-0.429†
	(1.404)	(0.383)	(0.480)	(0.230)
Second Treatment x Post	3.008†	-1.123*	1.714†	-1.279*
	(1.593)	(0.482)	(0.917)	(0.484)
Third Treatment x Post	6.016***	-1.083*	4.255**	-1.663**
	(1.579)	(0.475)	(1.365)	(0.581)
Firm FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
R^2	0.446	0.420	0.525	0.372
Adjusted R^2	0.431	0.404	0.512	0.355
Firms	38	38	38	38
Observations	20,520	20,520	20,520	20,520
Level	Firm-Minute	Firm-Minute	Firm-Minute	Firm-Minute

Table C.13: **Primary Field Study: Regression Analysis of Firm Productivity.** Using the same estimation method for firm process used in the main paper, we estimate the effect of our experimental treatment on the general productivity of the firms, measured in terms of *Lines*, representing the overall lines of software code written by the team, and ln(Lines + 1), the natural log of the same variable. Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	(1) Lines	(2) ln(Lines + 1)
Treatment x Post	31155.012	-0.403
	(36138.152)	(0.791)
Firm FE	Yes	Yes
Time FE	Yes	Yes
R^2	0.0993	0.592
Adjusted R^2	0.0750	0.581
Firms	38	38
Observations	20,520	20,520
Level	Firm-Minute	Firm-Minute

Table C.14: **Primary Field Study: Regression Analysis of Firm Process Controlling for Firm Productivity.** This table shows the same regression models for firm process as in the main paper, except including *ln(Lines + 1)* as a control for time-variant heterogeneity in firm productivity in generating software code, measured in terms of the logarithm of the lines of software code written by the firm. Ordinary least squares (OLS) estimation of firm-minute level data. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	(1) System-Level Branching	(2) Subsystem-Level Branching	(3) Code Integration Action	(4) Advanced API Specialization
Treatment x Post	3.653**	-0.892*	2.163*	-1.072**
	(1.139)	(0.382)	(0.845)	(0.386)
Ln(Lines + 1)	0.980***	0.222***	0.220**	0.128**
	(0.174)	(0.058)	(0.071)	(0.043)
Firm FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
R^2	0.560	0.541	0.481	0.381
Adjusted R^2	0.549	0.529	0.467	0.364
Firms	38	38	38	38
Observations	20,520	20,520	20,520	20,520
Level	Firm-Minute	Firm-Minute	Firm-Minute	Firm-Minute

Table C.15: **Primary Field Study: Comparison of Meeting Duration & Post-Meeting Latency.** The rows correspond to the sum of meeting duration and post-meeting duration for the three mentor meetings, with a stand-up meeting in treatment, and the total across all the meetings. The first three columns show the mean and standard deviation (in parentheses) for control, treatment, and the full sample. The last column shows the differences between treatment and control samples and the associated standard error (in parentheses). A t-test of difference in means finds no statistically significant difference between the treatment and control samples for any meeting or the total.

Meeting Duration & Post-Meeting Latency	Sample			Difference
	Control	Treatment	Full	
Meeting 1	13.70	19.00	16.21	5.300
	(23.56)	(23.89)	(23.55)	(7.706)
Meeting 2	14.50	9.944	12.34	-4.556
	(30.01)	(24.47)	(27.26)	(8.947)
Meeting 3	19.30	20.06	19.66	0.756
	(30.85)	(35.24)	(32.55)	(10.72)
Total for All Meetings	47.50	49.00	48.21	1.500
	(51.89)	(56.77)	(53.52)	(17.63)

Table C.16: **Primary Field Study: Regression Analysis of Firm Process Interacted with Firm Characteristics.** Ordinary least squares (OLS) estimation of firm-minute level data. We introduce firm characteristics of *Firm Size* and *Graduate Degree* as interaction terms; their baseline values drop out of the regression because they are time-invariant and collinear with the firm fixed effects. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	System-Level Branching			Subsystem-Level Branching		
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Treatment × Post	2.729	2.981†	2.059	-2.116*	-1.175**	-2.740*
	(3.048)	(1.724)	(3.560)	(0.914)	(0.421)	(1.041)
Treatment × Post	0.194		0.308	0.417		0.523
× Firm Size	(0.946)		(0.980)	(0.350)		(0.354)
Treatment × Post		0.737	0.961		0.516	0.895
× Graduate Degree		(2.927)	(3.037)		(0.520)	(0.538)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.410	0.410	0.411	0.425	0.418	0.431
Adjusted R^2	0.394	0.394	0.395	0.409	0.402	0.415
Firms	38	38	38	38	38	38
Observations	20,520	20,520	20,520	20,520	20,520	20,520
Level	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.

	Code Integration Action			Advanced API Specialization		
Variable	(7)	(8)	(9)	(10)	(11)	(12)
Treatment × Post	-5.454*	2.652*	-5.813*	-1.836*	-0.717	-1.154
	(2.126)	(1.175)	(2.626)	(0.840)	(0.524)	(0.862)
Treatment × Post	2.766**		2.826**	0.262		0.146
× Firm Size	(0.900)		(0.940)	(0.306)		(0.276)
Treatment × Post		-1.539	0.515		-1.084†	-0.978†
× Graduate Degree		(1.901)	(1.717)		(0.595)	(0.558)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.535	0.460	0.536	0.339	0.346	0.347
Adjusted R^2	0.523	0.445	0.523	0.321	0.328	0.329
Firms	38	38	38	38	38	38
Observations	20,520	20,520	20,520	20,520	20,520	20,520
Level	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.	Firm-Min.

Table C.17: **Primary Field Study: Mediation Analysis.** Generalized structural equation model (SEM) estimation on cross-sectional data at the firm level. All models are estimated in one structural model. *Code Integration Action* and *Advanced API Specialization* serve as mediators for both $Value$ and $Novelty$. Robust standard errors clustered at the firm level shown in parentheses, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Variable	Code Integration Action	Advanced API Specialization	Value	Novelty
Treatment	5.312***	-1.898***	-0.277	0.506
	(1.410)	(0.629)	(0.465)	(0.499)
Code Integration Action			0.147***	-0.030
			(0.047)	(0.031)
Advanced API Specialization			-0.093	0.404***
			(0.081)	(0.122)
Constant	1.828	1.043	3.906***	1.351
	(2.350)	(1.637)	(1.191)	(0.908)
Control Variables	Yes	Yes	Yes	Yes
Observations	38	38	38	38

Table C.18: **Follow-On Laboratory Study: Sample Size Required for Firm Outcomes and Process.** This a priori analysis was conducted under a mean difference t-test between two independent groups. The first two columns contain the estimated effect size of teams in Conditions 1 vs. 2 and Conditions 2 vs. 3, respectively. The last two columns list the computed required sample size in each comparison condition, given the different hypothesized α and power in parentheses.

	Effect Size		Sample Size	
	Conditions 1 vs. 2	Conditions 2 vs. 3	Conditions 1 vs. 2	Conditions 2 vs. 3
Outcomes				
Value	0.84	0.62	28 (10%, 80%)	50 (10%, 80%)
			32 (10%, 85%)	58 (10%, 85%)
			40 (10%, 90%)	70 (10%, 90%)
Novelty	0.51	0.69	72 (10%, 80%)	40 (10%, 80%)
			84 (10%, 85%)	48 (10%, 85%)
			104 (10%, 90%)	58 (10%, 90%)
Process				
Time to Integrate	0.61	0.60	50 (10%, 80%)	52 (10%, 80%)
			60 (10%, 85%)	62 (10%, 85%)
			72 (10%, 90%)	74 (10%, 90%)
Individual Sketches	0.76	0.96	34 (10%, 80%)	22 (10%, 80%)
			40 (10%, 85%)	26 (10%, 85%)
			48 (10%, 90%)	30 (10%, 90%)

Table C.19: **Follow-On Laboratory Study: Background Characteristics of Study Participants.** This table presents the means, standard deviations (in parentheses), and the F-statistic from an Analysis of Variance (ANOVA) for the background characteristics of the members of the teams. p-values for the F-test are shown in brackets. We do not find a statistically significant difference across these three groups in any of these background characteristics, verifying the validity of the experimental randomization. We average across the members of each team to form one observation per team. *Age* is measured in years. *Gender* indicates whether the subject is female (1) or male (0). *Graduate Education* indicates whether the individual has any graduate education. *Current Student* indicates whether the individual is a current student. *Any Experience* indicates whether the individual has any past experience in product development.

	Sample				
Variable	Condition 1	Condition 2	Condition 3	Total	F-Statistic
Age	35.35	35.86	33.67	34.97	0.627
	(7.567)	(6.836)	(6.561)	(6.960)	[0.537]
Gender	0.464	0.458	0.507	0.476	0.190
	(0.297)	(0.292)	(0.299)	(0.293)	[0.827]
Graduate Education	0.478	0.486	0.435	0.467	0.172
	(0.331)	(0.326)	(0.309)	(0.318)	[0.843]
Current Student	0.348	0.431	0.333	0.371	0.786
	(0.309)	(0.269)	(0.284)	(0.287)	[0.460]
Any Experience	0.203	0.236	0.145	0.195	0.682
	(0.280)	(0.269)	(0.263)	(0.269)	[0.509]

Table C.20: **Follow-On Laboratory Study: Additional Variable Definitions and Sources.** Additional measures of outcomes, process, intervention duration, and ad hoc communication drawn from each team's *Final Output* design, *Video Recording* of their working session, their *Individual Sketches*, and a *Post Survey* after the experiment.

Variable	Definition	Source
Outcomes		
Completeness	Completeness of the final product (Likert 1–5). Average of two independent rater assessments.	Final Output
Process		
Coordination (Multiple)	Self-reported responses to five survey measures of coordination performance based on Lewis (2003) (Likert 1–5): *Effectiveness*, *Few Misunderstandings*, *Low Backtracking*, *Efficiency*, and *Low Confusion*.	Post Survey
Specialization (Multiple)	Self-reported responses to five survey measures of individual specialization performance based on Lewis (2003) (Likert 1–5): *Group*, *Individual*, *Responsibility*, *Necessity*, and *Awareness*.	Post Survey
Intervention Duration		
Meeting Duration	Duration in seconds of all formal meetings associated with experimental intervention(s), measured from the time the mentor enters the room to when he exits, which is after the last statement by any member in response to a intervention question.	Video Recording
Post-Meeting Latency	Duration in seconds of the total time after the mentor exits the room and until the team begins working on either individual sketches or engages in ad hoc communication related to the product.	Video Recording
Ad Hoc Communication		
Ad Hoc Frequency	Count of distinct exchanges not including those related to the experimental intervention, where each exchange is an individual's oral communication bounded by speech by other individuals on the team.	Video Recording
Ad Hoc Word Count	Count of words spoken by anyone on the team not including those related to the experimental intervention.	Video Recording
Deadline Salience		
Time	Proportion of words on matters related to time, e.g., end, until.	Video Recording
Future Focus	Proportion of words that suggest a focus on the future, e.g., may, will, soon	Video Recording
Discrepancy	Proportion of words that are verbs indicating thinking or action that are different from the status quo, e.g., should would.	Video Recording
Anxiety	Proportion of words that indicate anxiety, e.g., worried, fearful.	Video Recording
Negative Emotion	Proportion of words that indicate negative affective, e.g., hurt, ugly, nasty.	Video Recording
Swear	Proportion of words that are informal and vulgar terms, e.g., damn.	Video Recording

Table C.21: **Follow-On Laboratory Study: Additional Variables Summary Statistics and Cross-Sectional Analysis.** The first three columns contain the mean and in parentheses the standard deviation of teams in each condition. The last two columns compare Conditions 1 vs. 2 and Conditions 2 vs. 3, respectively, based on a t-test of the difference in means; the values reflect the difference in means and in parentheses the standard error, with significance indicated by $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

	Sample				Difference in Means	
	Full	Condition 1	Condition 2	Condition 3	1 vs. 2	2 vs. 3
Outcomes						
Completeness	3.381	3.362	3.389	3.391	0.0266	0.00242
	(0.675)	(0.778)	(0.753)	(0.478)	(0.223)	(0.185)
Process						
Coor.: Effectiveness	4.133	3.580	4.181	4.638	0.601**	0.457**
	(0.719)	(0.760)	(0.461)	(0.481)	(0.182)	(0.137)
Coor.: Few Misund.	4.062	3.623	4.083	4.478	0.460†	0.395†
	(0.872)	(0.895)	(0.806)	(0.724)	(0.248)	(0.224)
Coor.: Low Bktrk.	2.705	2.319	2.694	3.101	0.376	0.407†
	(0.800)	(0.693)	(0.839)	(0.685)	(0.225)	(0.224)
Coor.: Efficiency	4.095	3.913	4.153	4.217	0.240	0.0646
	(0.817)	(1.065)	(0.674)	(0.656)	(0.259)	(0.194)
Coor.: Low Confsn.	2.862	2.638	2.764	3.188	0.126	0.425†
	(0.734)	(0.643)	(0.813)	(0.642)	(0.214)	(0.214)
Spec.: Group	3.733	4.000	3.722	3.478	-0.278†	-0.244
	(0.656)	(0.522)	(0.570)	(0.771)	(0.160)	(0.197)
Spec.: Individual	3.248	3.754	3.264	2.725	-0.490*	-0.539*
	(0.891)	(0.812)	(0.792)	(0.789)	(0.234)	(0.231)
Spec.: Responsibility	3.271	3.333	3.278	3.203	-0.0556	-0.0749
	(0.736)	(0.876)	(0.570)	(0.764)	(0.215)	(0.196)
Spec.: Necessity	3.776	3.841	3.778	3.710	-0.0628	-0.0676
	(0.780)	(0.828)	(0.727)	(0.812)	(0.227)	(0.225)
Spec.: Awareness	3.414	3.406	3.361	3.478	-0.0447	0.117
	(0.726)	(0.531)	(0.589)	(0.999)	(0.164)	(0.238)
Intervention Duration						
Meeting Duration	101.0	68.83	98.63	135.6	29.80*	36.98*
	(53.10)	(36.94)	(50.46)	(49.76)	(12.95)	(14.63)
Post-Meeting Latency	38.61	32.74	31.25	52.17	-1.489	20.92*
	(30.42)	(26.46)	(26.70)	(34.19)	(7.757)	(8.927)
Ad Hoc Communication						
Ad Hoc Frequency	162.2	127.4	161.8	197.3	34.44†	35.43†
	(68.17)	(52.79)	(78.39)	(53.02)	(19.58)	(19.61)
Ad Hoc Word Count	5749.8	5611.3	5795	5841.1	183.7	46.09
	(1774.7)	(1630.0)	(1833.7)	(1917.1)	(506.9)	(547.1)
Deadline Salience						
Time	3.358	2.997	3.332	3.745	0.334*	0.413*
	(0.601)	(0.480)	(0.391)	(0.671)	(0.128)	(0.159)
Future Focus	1.469	1.517	1.408	1.486	-0.109	0.0780
	(0.310)	(0.275)	(0.324)	(0.331)	(0.0878)	(0.0955)
Discrepancy	2.815	2.391	2.829	3.223	0.437**	0.394†
	(0.698)	(0.467)	(0.523)	(0.813)	(0.145)	(0.199)
Anxiety	0.0862	0.0883	0.0892	0.0810	0.000906	-0.00820
	(0.0594)	(0.0651)	(0.0625)	(0.0520)	(0.0186)	(0.0168)
Negative Emotion	0.652	0.636	0.675	0.644	0.0393	-0.0307
	(0.187)	(0.187)	(0.199)	(0.180)	(0.0564)	(0.0555)
Swear	0.0774	0.0726	0.0904	0.0687	0.0178	-0.0217
	(0.0584)	(0.0670)	(0.0544)	(0.0531)	(0.0178)	(0.0157)

Table C.22: **Follow-On Laboratory Study: Survey-Based Measures of Coordination and Specialization.** These survey measures come from Lewis (2003) and subsequent work.

Variable	Survey Question
Coordination	
Effectiveness	"Our team worked together in a well-coordinated fashion."
Few Misunderstandings	"Our team had very few misunderstandings about what to do."
Low Backtracking	"Our team needed to backtrack and start over a lot." (Reversed)
Efficiency	"We accomplished the task smoothly and efficiently."
Low Confusion	"There was much confusion about how we would accomplish the task." (Reversed)
Specialization	
Group	"Each team member has specialized knowledge of some aspect of our project."
Individual	"I have knowledge about an aspect of the project that no other team member has."
Responsibility	"Different team members are responsible for expertise in different areas."
Necessity	"The specialized knowledge of several different team members was needed to complete the project deliverables."
Awareness	"I know which team members have expertise in specific areas."

Table C.23: **Follow-On Laboratory Study: Correlations.** This table shows pairwise correlations for the dependent variables used in the follow-on laboratory study.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	*Value*	1									
(2)	*Novelty*	0.018	1								
(3)	*Completeness*	0.287	0.460	1							
(4)	*Time to Integrate*	-0.140	0.169	-0.012	1						
(5)	*Individual Sketches*	-0.274	0.151	0.065	0.483	1					
(6)	*Coor.: Effectiveness*	0.352	-0.194	0.053	-0.407	-0.280	1				
(7)	*Coor.: Few Misund.*	0.395	-0.105	0.102	-0.389	-0.145	0.734	1			
(8)	*Coor.: Low Bktrk.*	0.370	-0.156	0.110	-0.056	-0.135	0.563	0.474	1		
(9)	*Coor.: Efficiency*	0.114	-0.158	0.044	-0.154	0.066	0.699	0.654	0.495	1	
(10)	*Coor.: Low Confsn.*	0.326	0.002	0.208	-0.092	-0.054	0.551	0.486	0.793	0.438	1
(11)	*Spec.: Group*	-0.311	-0.046	-0.102	0.125	0.227	0.097	0.159	0.041	0.274	0.059
(12)	*Spec.: Individual*	-0.429	0.045	-0.181	0.067	0.203	-0.284	-0.140	-0.355	-0.057	-0.223
(13)	*Spec.: Responsibility*	-0.192	-0.090	-0.120	-0.083	-0.111	0.135	0.089	-0.130	0.141	-0.123
(14)	*Spec.: Necessity*	-0.177	-0.081	-0.059	-0.033	0.020	0.255	0.141	-0.043	0.120	-0.066
(15)	*Spec.: Awareness*	-0.168	-0.150	-0.130	-0.037	-0.194	0.241	0.251	0.128	0.204	0.130
(16)	*Meeting Duration*	0.447	-0.105	0.052	-0.069	-0.390	0.213	0.204	0.190	-0.103	0.071
(17)	*Post-Meeting Latency*	0.174	-0.033	0.017	-0.008	-0.162	0.225	0.110	0.127	0.013	0.037
(18)	*Ad Hoc Frequency*	0.175	-0.012	0.076	-0.030	-0.166	0.263	0.257	0.212	0.072	0.196
(19)	*Ad Hoc Word Count*	-0.062	0.084	-0.055	0.069	-0.037	-0.157	-0.206	0.004	-0.358	-0.052
(20)	*Time*	0.322	-0.294	-0.120	-0.210	-0.251	0.323	0.262	0.273	0.081	0.274
(21)	*Future Focus*	-0.033	0.101	0.080	-0.120	-0.049	0.017	0.039	0.084	0.019	0.049
(22)	*Discrepancy*	0.310	-0.067	0.276	-0.222	-0.310	0.130	-0.034	0.057	-0.195	0.071
(23)	*Anxiety*	-0.034	0.165	-0.076	-0.093	-0.071	-0.084	-0.072	0.026	-0.250	0.105
(24)	*Negative Emotion*	-0.121	0.024	-0.123	0.001	-0.065	-0.039	-0.232	-0.042	-0.179	-0.038
(25)	*Swear*	-0.013	-0.107	-0.124	0.068	-0.093	0.151	0.021	0.021	0.181	-0.032

		(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(11)	*Spec.: Group*	1								
(12)	*Spec.: Individual*	0.420	1							
(13)	*Spec.: Responsibility*	0.602	0.456	1						
(14)	*Spec.: Necessity*	0.540	0.301	0.596	1					
(15)	*Spec.: Awareness*	0.625	0.419	0.582	0.525	1				
(16)	*Meeting Duration*	-0.240	-0.289	-0.045	-0.143	0.195	1			
(17)	*Post-Meeting Latency*	-0.171	-0.307	-0.030	-0.011	0.003	0.449	1		
(18)	*Ad Hoc Frequency*	-0.097	-0.255	0.068	-0.082	0.095	0.413	0.044	1	
(19)	*Ad Hoc Word Count*	-0.333	-0.038	-0.212	-0.092	-0.142	0.193	0.030	0.249	1
(20)	*Time*	-0.251	-0.417	-0.183	-0.189	-0.130	0.247	0.173	0.079	-0.189
(21)	*Future Focus*	-0.053	-0.163	0.065	0.117	0.076	0.114	0.042	-0.077	0.011
(22)	*Discrepancy*	-0.392	-0.250	-0.089	-0.071	-0.200	0.180	-0.019	0.258	0.233
(23)	*Anxiety*	-0.121	0.085	-0.054	-0.053	-0.045	-0.012	0.074	-0.096	0.269
(24)	*Negative Emotion*	-0.234	-0.199	-0.173	-0.095	-0.198	0.021	0.098	-0.068	0.202
(25)	*Swear*	-0.017	-0.182	-0.051	0.027	-0.067	-0.087	0.048	-0.110	-0.121

		(20)	(21)	(22)	(23)	(24)	(25)
(20)	*Time*	1					
(21)	*Future Focus*	0.164	1				
(22)	*Discrepancy*	-0.026	-0.017	1			
(23)	*Anxiety*	-0.026	-0.066	0.069	1		
(24)	*Negative Emotion*	-0.014	-0.133	0.030	0.502	1	
(25)	*Swear*	-0.036	-0.123	-0.182	-0.140	0.297	1

Bibliography

Adner R, Levinthal DA (2004) What is not a real option: Considering boundaries for the application of real options to business strategy. *Academy of Management Review* 29(1):74–85.

Aggarwal VA, Hsu DH, Wu A (2020) Organizing knowledge production teams within firms for innovation. *Strategy Science* 5(1):1–16.

Agrawal A, Gans JS, Stern S (2021) Enabling Entrepreneurial Choice. *Management Science* forthcomin.

Ahuja G, Morris Lampert C (2001) Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal* 22(6–7):521–543.

Albert D (2018) Organizational module design and architectural inertia: Evidence from structural recombination of business decisions. *Organization Science* 29(5):890–911.

Aldrich HE (1979) *Organizations and Environments* (Englewood Cliffs, NJ: Prentice-Hall).

Alexander L, Van Knippenberg D (2014) Teams in pursuit of radical innovation: A goal orientation perspective. *Academy of Management Review* 39(4):423–438.

Alvarez SA, Barney JB (2010) Entrepreneurship and epistemology: The philosophical underpinnings of the study of entrepreneurial opportunities. *Academy of Management Annals* 4(1):557–583.

Amabile T (1996) *Creativity in Context: Update to the Social Psychology of Creativity* (Boulder, CO: Westview Press).

Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* 45(2):357–376.

Amabile TM, Pratt MG (2016) The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior* 36:157–183.

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

Anderson C, Galinsky A (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.

Angrist JD, Pischke JS (2008) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Ante SE (2011) Skype to acquire start-up GroupMe. *Wall Street Journal* .

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

Arrington M (2010) GroupMe, born at TechCrunch Disrupt, secures funding and launches. *TechCrunch* .

Ashforth BE, Mael F (1989) Social identity theory and the organization. *Academy of Management Review* 14(1):20–39.

Azevedo EM, Deng A, Luis J, Olea M, Rao J, Weyl EG (2019) A/B Testing with Fat Tails.

Bagozzi RP, Yi Y, Phillips LW (1991) Assessing construct validity in organizational research. *Administrative Science Quarterly* 36(3):421–458.

Baker T, Nelson RE (2005) Creating Something from Nothing: Resource Construction through Entrepreneurial Bricolage. *Administrative Science Quarterly* 50:329–366.

Baldwin C, Clark KB (2000) *Design Rules: The Power of Modularity* (Cambridge, MA: MIT Press).

Baron RA (2004) The cognitive perspective: A valuable tool for answering entrepreneurship's basic "why" questions. *Journal of Business Venturing* 19(2):221–239.

Baudet GM (1978) On the branching factor of the alpha-beta pruning algorithm. *Artificial Intelligence* 10(2):173–199.

Baumann O, Schmidt J, Stieglitz N (2019) Effective Search in Rugged Performance Landscapes: A Review and Outlook. *Journal of Management* 45(1):285–318.

Beck K, Beedle M, Van Bennekum A, Cockburn A, Cunningham W, Fowler M, Grenning J, Highsmith J, Hunt A, Jeffries R, Kern J, Marick B, Martin RC, Mellor S, Schwaber K, Sutherland J, Thomas D (2001) Manifesto for Agile software development .

Bennett VM, Chatterji AK (2019) The entrepreneurial process: Evidence from a nationally representative survey. *Strategic Management Journal* (June):1–31.

Berg JM (2014) The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes* 125(1):1–17.

Bernstein E, Shore J, Lazer D (2019) Improving the rhythm of your collaboration. *MIT Sloan Management Review* 61(1):29–36.

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119(1):249–275.

Besemer S (1998) Creative product analysis matrix: Testing the model structure and a comparison among products–three novel chairs. *Creativity Research Journal* 11(4):333–346.

Beshears J, Choi JJ, Laibson D, Madrian BC, Milkman KL (2015) The Effect of Providing Peer Information on Retirement Savings Decisions. *Journal of Finance* 70(3):1161–1201.

Billinger S, Stieglitz N, Schumacher TR (2014) Search on rugged landscapes: An experimental study. *Organization Science* 25(1):93–108.

Bingham CB, Davis JP (2012) Learning sequences: Their existence, effect, and evolution. *Academy of Management Journal* 55(3):611–641.

Birkinshaw J (2018) What to expect from Agile. *MIT Sloan Management Review* 59(2):39–42.

Blank S (2003) *The four steps to the epiphany: successful strategies for products that win* (John Wiley & Sons).

Bocken N, Snihur Y (2020) Lean Startup and the business model: Experimenting for novelty and impact. *Long Range Planning* 53(4):101953.

Bohn R, Lapré MA (2011) Accelerated learning by experimentation. Jaber MY, ed., *Learning Curves: Theory, Models, and Applications*, 191–209 (Boca Raton, FL: CRC Press), ISBN 9781439807408, URL `http://dx.doi.org/10.2139/ssrn.1640767`.

Broussard M (2015) The secret lives of hackathon junkies. *The Atlantic* .

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* 39:24–24.

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

Burton RM, Håkonsson DD, Nickerson J, Puranam P, Workiewicz M, Zenger T (2017) GitHub: exploring the space between boss-less and hierarchical forms of organizing. *Journal of Organization Design* 6(1):10.

Byrne BM (2001) *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming* (Mahwah, NJ: Lawrence Erlbaum Associates).

Byron K, Khazanchi S (2012) Rewards and creative performance: A meta-analytic test of theoretically derived hypotheses. *Psychological Bulletin* 138(4):809–830.

Camuffo A, Cordova A, Gambardella A, Spina C (2020a) A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial. *Management Science* 66(2):564–586.

Camuffo A, Gambardella A, Spina C (2020b) Small changes with big impact: Experimental evidence of a scientific approach to the decision-making of entrepreneurial firms .

Cao L, Ramesh B (2007) Agile software development: Ad hoc practices or sound principles? *IT Professional* 9(2):41–47.

Carson PP, Carson K (1993) Managing creativity enhancement through goal-setting and feedback. *Journal of Creative Behavior* 27(1):36–45.

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

Chatterji AK, Delecourt S, Hasan S, Koning R (2019) When does advice impact startup performance? *Strategic Management Journal* 40(3):331–356.

Chatterji AK, Findley M, Jensen NM, Meier S, Nielson D (2016) Field experiments in strategy research. *Strategic Management Journal* 37(1):116–132.

Christensen CM, Bower JL (1996) Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal* 17(3):197–218.

Clough DR, Fang TP, Bala Vissa B, Wu A (2019) Turning lead into gold: How do entrepreneurs mobilize resources to exploit opportunities? *Academy of Management Annals* 13(1):240–271.

Contigiani A, Levinthal DA (2019) Situating the construct of lean start-up: Adjacent conversations and possible future directions. *Industrial and Corporate Change* 28(3):551–564.

Critcher CR, Ferguson MJ (2016) "Whether I like it or not, it's important": Implicit importance of means predicts self-regulatory persistence and success. *Journal of Personality and Social Psychology* 110(6):818–839.

Cromwell JR, Amabile TM, Harvey JF (2018a) An integrated model of dynamic problem solving within organizational constraints. *Individual Creativity in the Workplace*, 53–81 (Elsevier).

Cromwell JR, Amabile TM, Harvey JF (2018b) An integrated model of dynamic problem solving within organizational constraints. Reiter-Palmon R, Kennel VL, Kaufman JC, eds., *Individual Creativity in the Workplace*, volume 7 of *Explorations in Creativity Research*, 53–81 (London: Academic Press).

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

Csaszar FA (2018) What Makes a Decision Strategic? Strategic Representations. *Strategy Science* 3(4):606–619.

Csaszar FA, Laureiro-Martínez D (2018) Individual and Organizational Antecedents of Strategic Foresight. *Strategy Science* 3(3):513–532.

Csaszar FA, Levinthal DA (2016) Mental representation and the discovery of new strategies. *Strategic Management Journal* 37:2031–2049.

Csikszentmihalyi M (1996) *Creativity: Flow and the Psychology of Discovery and Invention* (New York: HarperCollins).

Cyert RM, March JG (1963) *A Behavioral Theory of the Firm* (Englewood Cliffs, NJ: Prentice Hall).

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

Dalton, DR, Todor, WD, Spendolin, MJ, Fielding, GJ and Potor L (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. Mumford M, ed., *Handbook of Organizational Creativity*, 483–513 (Oxford, UK: Elsevier).

Delmar F, Shane S (2003) Does business planning facilitate the development of new ventures? *Strategic Management Journal* 24(12):1165–1185.

Deng A, Xu Y, Kohavi R, Walker T (2013) Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining* 123–132.

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.

Dobrajska M, Billinger S, Karim S (2015) Delegation within hierarchies: How information processing and knowledge characteristics influence the allocation of formal and real decision authority. *Organization Science* 26(3):687–704.

Duflo E, Saez E (2003) The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Quarterly Journal of Economics* 118(3):815–842.

Edelman B, Lai Z (2016) Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research* 53(6):881–900.

Edmondson AC, McManus SE (2007) Methodological fit in field research. *Academy of Management Review* 32(4):1155–1179.

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.

Eggers JP, Kaul A (2018) Motivation and ability? A behavioral perspective on the pursuit of radical invention in multi-technology incumbents. *Academy of Management Journal* 61(1):67–93.

Eisenberger R, Rhoades L (2001) Incremental effects of reward on creativity. *Journal of Personality and Social Psychology* 81(4):728–741.

Eisenhardt KM, Bingham CB (2017) Superior strategy in entrepreneurial settings: Thinking, doing, and the logic of opportunity. *Strategy Science* 2(4):246–257.

Eisenhardt KM, Tabrizi BN (1995) Accelerating adaptive processes: Product innovation in the global computer industry. *Administrative Science Quarterly* 40(1):84–110.

Eisenmann T, Ries E, Dillard S (2013) Hypothesis-Driven Entrepreneurship: The Lean

Startup. *Background note 812-095, Harvard Business School*, HBS (Cambridge, MA), ISBN 8005457685.

Eric E, Spencer S, Stricchiola J (2015) The art of seo: Mastering search engine optimization.

Ethiraj SK, Levinthal D (2009) Hoping for A to Z while rewarding only A: Complex organizations and multiple goals. *Organization Science* 20(1):4–21.

Ethiraj SK, Levinthal DA (2004) Bounded rationality and the search for organizational architecture: An evolutionary perspective on the design of organizations and their evolvability. *Administrative Science Quarterly* 49(3):404–437.

Ewens M, Nanda R, Rhodes-Kropf M (2018) Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics* 128(3):422–442.

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

Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39(2):175–191.

Felin T, Gambardella A, Stern S, Zenger T (2019) Lean startup and the business model: Experimentation revisited. *Long Range Planning* (June).

Felin T, Zenger TR (2009) Entrepreneurs as Theorists: On the Origins of Collective Beliefs and Novel Strategies. *Strategic Entrepreneurship Journal* 3:127–146.

Fleming L (2001) Recombinant uncertainty in technological search. *Management Science* 47(1):117–132.

Fornell C, Larcker DF (1981a) Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18(1):39–50.

Fornell C, Larcker DF (1981b) Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research* 18(1):382–388.

Furr N, O'Keeffe K, Dyer JH (2016) Managing multiparty innovation. *Harvard Business Review* 94(11):76–83.

Gaba V, Greve HR (2019) Safe or profitable? The pursuit of conflicting goals. *Organization Science* 30(4):647–667.

Gabe G (2018) Trapped In Google's New Video Carousels – A Dangerous SERP Feature For Some Ecommerce Retailers [Case Study]. URL `https://www.gsqi.com/marketing-blog/trapped-in-google-video-carousels-case-study/`.

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

Ganz SC (2020) Hyperopic Search: Organizations Learning About Managers Learning About Strategies. *Organization Science* (April).

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

Gary MS, Wood RE (2011) Mental Models, Decision Rules, and Performance Heterogenity. *Strategic Management Journal* 32(6):569–594.

Gavetti G, Greve HR, Levinthal DA, Ocasio W (2012) The behavioral theory of the firm: Assessments and prospects. *Academy of Management Annals* 6(1):1–40.

Gavetti G, Levinthal DA (2000) Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly* 45(1):113–137.

Gavetti G, Levinthal DA, Rivkin JW (2005) Strategy making in novel and complex worlds: The power of analogy. *Strategic Management Journal* 26(8):691–712.

Gavetti G, Menon A (2016) Evolution Cum Agency: Toward a Model of Strategic Foresight. *Strategy Science* 1(3):207–233.

Gavetti G, Rivkin JW (2007) On the Origin of Strategy: Action and Cognition over Time. *Organization Science* 18(3):420–439.

Gersick C (1988) Time and transition in work teams: Toward a new model of group development. *Academy of Management Journal* 31(1):9–41.

Gersick C (1989) Marking time: Predictable transitions in task groups. *Academy of Management Journal* 32(2):274–309.

Ghemawat P (1991) *Commitment: The Dynamic of Strategy* (New York, NY: Free Press).

Ghosh S (2020) Think before you act: The unintended consequences of inexpensive business experimentation. *Harvard Business School Working Paper* .

Ghosh S, Thomke SH, Pourkhalkhali H (2020) The effects of hierarchy on learning and performance in business experimentation. *Harvard Business School Working Paper* .

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management Science* 56(4):591–605.

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.

Grant RM (1996) Toward a knowledge-based theory of the firm. *Strategic Management Journal* 17(S2):109–122.

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.

Greve HR, Gaba V (2020) Performance feedback in organizaitons and groups: Common themes. Argote L, Levine JM, eds., *The Handbook of Group and Organizational Learning* (Oxford, UK: Oxford University Press).

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

Gupta S, Kohavi R, Tang D, Xu Y, Andersen R, Bakshy E, Cardin N, Chandran S, Chen N, Coey D, Curtis M, Deng A, Duan W, Forbes P, Frasca B, Guy T, Imbens GW, Saint Jacques G, Kantawala P, Katsev I, Katzwer M, Konutgan M, Kunakova E, Lee M, Lee M, Liu J, McQueen J, Najmi A, Smith B, Trehan V, Vermeer L, Walker T, Wong J, Yashkov I (2019) Top Challenges from the first Practical Online Controlled Experiments Summit. *ACM SIGKDD Explorations Newsletter* 21(1):20–35.

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

Hall R, Tolbert P (2005) *Organizations: Structures, processes, and outcomes* (Upper Saddle River, NJ: Prentice Hall), 8th edition.

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.

Helfat CE, Peteraf MA (2015) Managerial Cognitive Capabilities and the Microfoundations of Dynamic Capabilities. *Strategic Management Journal* 36:831–850.

Henderson RM, Clark KB (1990) Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly* 35(1):9–30.

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

Hoppmann J, Naegele F, Girod B (2019) Boards as a source of inertia: Examining the internal challenges and dynamics of boards of directors in times of environmental discontinuities. *Academy of Management Journal* 62(2):437–468.

Hu S, Bettis RA (2018) Multiple organization goals with feedback from shared technological task environments. *Organization Science* 29(5):873–889.

Huang L, Pearce JL (2015) *Managing the Unknowable: The Effectiveness of Early-stage Investor Gut Feel in Entrepreneurial Investment Decisions*, volume 60.

Im S, Workman JP (2004) Market orientation, creativity, and new product performance in high-technology firms. *Journal of Marketing* 68(2):114–132.

Jick TD (1979) Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly* 24(4):602–611.

Johnson G, Shriver S, Goldberg S (2020) Privacy & market concentration: Intended & unintended consequences of the gdpr. *Available at SSRN 3477686* .

Joseph J, Gaba V (2020) Organizational structure, information processing, and decision-making: A retrospective and road map for research. *Academy of Management Annals* 14(1):267–302.

Joseph J, Klingebiel R, Wilson AJ (2016) Organizational structure and performance feedback: Centralization, aspirations, and termination decisions. *Organization Science* 27(5):1065–1083.

Jung HJ, Lee JJ (2016) The quest for originality: A new typology of knowledge search and breakthrough inventions. *Academy of Management Journal* 59(5):1725–1753.

Kacperczyk A, Marx M (2016) Revisiting the small-firm effect on entrepreneurship: Evidence from firm dissolutions. *Organization Science* 27(4):893–910.

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

Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal* 1457(36):1435–1457.

Katila R, Shane S (2005) When Does Lack of Resources Make New Firms Innovative? *Academy of Management Journal* 48(5):814–829.

Kerr WR, Nanda R, Rhodes-Kropf M (2014) Entrepreneurship as experimentation. *Journal of Economic Perspectives* 28(3):25–48.

Keum DD, See E (2017) The influence of hierarchy on idea generation and selection in the innovation process. *Organization Science* 28(4):653–669.

Kirtley J, O'Mahony S (2020) What is a pivot? Explaining when and how entrepreneurial firms decide to make strategic change and pivot. *Strategic Management Journal* 1–34.

Klein PG (2008) Opportunity Discovery, Entrepreurial Action, and Economic Organization. *Strategic Entrepreneurship Journal* 2:175–190.

Knight K (1963) *A Study of Technological Innovation: The Evolution of Digital Computers*. Phd dissertation, Carnegie Institute of Technology.

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

Knudsen T, Srikanth K (2014) Coordinated exploration: Organizing joint search by multiple specialists to overcome mutual confusion and joint myopia. *Administrative Science Quarterly* 59(3):409–441.

Knuth DE, Moore RW (1975) An analysis of alpha-beta pruning. *Artificial Intelligence* 6(4):293–326.

Kohavi R, Henne RM, Sommerfield D (2007) Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. *Proceedings of The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2007, 959, ISBN 9781595936097, URL http://dx.doi.org/10.1145/1281192.1281295.

Kohavi R, Tang D, Xu Y (2020) *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (New York, NY: Cambridge University Press).

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. *Harvard Business School Working Paper No. 20-018* 1–46.

Lawrence PR, Lorsch JW (1967) Differentiation and integration in complex organizations. *Administrative Science Quarterly* 1(12):1–47.

Leatherbee M, Katila R (2019) The lean startup method: Team composition, hypothesis-testing, and early-stage business models. *Hypothesis-testing, and Early-stage Business Models (August 15, 2019)* .

Leatherbee M, Katila R (2020) The lean startup method: Early-stage teams and hypothesis-based probing of business ideas. *Strategic Entrepreneurship Journal* (September):1–24.

Leckart S (2012) The hackathon is on: Pitching and programming the next killer app. *Wired* .

Leckart S (2015) The hackathon fast track, from campus to Silicon Valley. *New York Times* .

Lee F, Edmondson AC, Thomke S, Worline M (2004) The mixed effects of inconsistency on experimentation in organizations. *Organization Science* 15(3):310–326.

Lee MY, Edmondson AC (2017) Self-managing organizations: Exploring the limits of less-hierarchical organizing. *Research in Organizational Behavior* 37:35–58.

Lee S (2020) The Myth of the Flat Start-up : Reconsidering the Organizational Structure of Start-ups.

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

Leiblein MJ, Reuer JJ, Zenger T (2018) What Makes a Decision Strategic? *Strategy Science* 3(4):558–573.

Levine SS, Prietula MJ (2014) Open collaboration for innovation: Principles and performance. *Organization Science* 25(5):1414–1433.

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.

Levinthal DA, Rerup C (2020) The plural of goal: Learning in a world of ambiguity. *Organization Science* (forthcoming).

Levinthal DA, Workiewicz M (2018) When two bosses are better than one: Nearly decomposable systems and organizational adaptation. *Organization Science* 29(2):207–224.

Levitt SD, List JA (2011) Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics* 3(1):224–238.

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.

Lewis K (2003) Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology* 88(4):587–604.

Lifshitz-Assaf H, Lebovitz S, Zalmanson L (2020) Minimal and adaptive coordination: How hackathons' projects accelerate innovation without killing it. *Academy of Management Journal* Forthcoming.

Lincoln JR, Miller J (1979) Work and friendship ties in organizations: A comparative analysis of relation networks. *Administrative Science Quarterly* 24(2):181–199.

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

Lu S, Bartol KM, Venkataramani V, Zheng X, Liu X (2019) Pitching novel ideas to the boss: The interactive effects of employees' idea enactment and influence tactics on creativity assessment and implementation. *Academy of Management Journal* 62(2):579–606.

March JG (1991) Exploration and exploitation in organizational learning. *Organization Science* 2(1):71–87.

March JG, Simon HA (1958) *Organizations* (New York: Wiley).

Masicampo EJ, Baumeister RF (2011) Consider it done! Plan making can eliminate the cognitive effects of unfulfilled goals. *Journal of Personality and Social Psychology* 101(4):667–683.

McCarthy M, Chen CC, McNamee RC (2018) Novelty and usefulness trade-off: Cultural cognitive differences and creative idea evaluation. *Journal of Cross-Cultural Psychology* 49(2):171–198.

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

McKenzie D (2012) Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics* 99(2):210–221.

Meyer MN, Heck PR, Holtzman GS, Anderson SM, Cai W, Watts DJ, Chabris CF (2019) Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences* 116(22):10723–10728.

Meyer MW, Zucker LG (1989) *Permanently Failing Organizations* (Newbury Park, CA: Sage).

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

Miner AS, Bassoff P, Moorman C (2001) Organizational and Improvisation Learning: A Field Study. *Administrative Science Quarterly* 46(2):304–337.

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

Miron-Spektor E, Beenen G (2015) Motivating creativity: The effects of sequential and simultaneous learning and performance achievement goals on product novelty and usefulness. *Organizational Behavior and Human Decision Processes* 127:53–65.

Montgomery DC (2013) *Design and Analysis of Experiments* (Wiley), 8th edition.

Mosakowski E (1997) Strategy Making under Causal Ambiguity: Conceptual Issues and Empirical Evidence. *Organization Science* 8(4):414–442.

Mueller JS, Melwani S, Goncalo JA (2012) The bias against creativity. *Psychological Science* 23(1):13–17.

Muja M, Lowe DG (2009) Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP 2009—Proceedings of the 4th International Conference on Computer Vision Theory and Applications* 2(2):331–340.

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

Nayak PR, Ketteringham J (1997) 3M's Post-it Notes: A Managed or Accidental Innovation? Katz R, ed., *The Human Side of Managing Technological Innovation*, 367–377 (Oxford University Press).

Nolte A, Pe-Than EPP, Filippova A, Bird C, Scallen S, Herbsleb JD (2018) You hacked and now what? - Exploring outcomes of a corporate hackathon. *Proceedings of the ACM Human-Computer Interaction* 2(CSCW).

Nonaka I (1994) A dynamic theory of organizational knowledge creation. *Organization Science* 5(1):14–37.

Obloj T, Sengul M (2020) What do multiple objectives really mean for performance? Empirical evidence from the French manufacturing sector. *Strategic Management Journal* 41(13):2518–2547.

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

Okhuysen G, Bechky BA (2009) Coordination in organizations: An integrative perspective. *Academy of Management Annals* 3(1):463–502.

Okhuysen G, Eisenhardt KM (2002) Integrating in groups: How formal knowledge interventions enable flexibility. *Organization Science* 13(4):370–386.

Oldham GR, Cummings A (1996) Employee creativity: Personal and contextual factors at work. *Academy of Management Journal* 39(3):607–634.

Ordóñez LD, Schweitzer ME, Galinsky AD, Bazerman MH (2009) Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives* 23(1):6–16.

Ott TE, Eisenhardt KM, Bingham CB (2017) Strategy formation in entrepreneurial settings: Past insights and future directions. *Strategic Entrepreneurship Journal* 11(3):306–325.

Packard MD, Clark BB, Klein PG (2017) Uncertainty Types and Transitions in the Entrepreneurial Process. *Organization Science* 28(5):840–856.

Paletz SB, Peng K (2008) Implicit theories of creativity across cultures: Novelty and appropriateness in two product domains. *Journal of Cross-Cultural Psychology* 39(3):286–302.

Pan Fang T, Wu A, Clough DR (2020) Platform diffusion at temporary gatherings: Social coordination and ecosystem emergence. *Strategic Management Journal* Forthcoming.

Parnas DL (1972) On the criteria to be used in decomposing systems into modules. *Communications of the ACM* 15(12):1053–1058.

Pe-Than EPP, Nolte A, Filippova A, Bird C, Scallen S, Herbsleb JD (2019) Designing corporate hackathons with a purpose: The future of software development. *IEEE Software* 36(1):15–22.

Pekelis L, Walsh D, Johari R (2015) The new stats engine .

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. Technical report.

Pervin LA (1992) The rational mind and the problem of volition. *Psychological Science* 3(3):162–164.

Pich MT, Loch CH, Meyer AD (2000) On Uncertainty , Ambiguity , and Complexity in Project Management 7–12.

Pillai SD, Goldfarb B, Kirsch DA (2020) The origins of firm strategy: Learning by economic experimentation and strategic pivots in the early automobile industry. *Strategic Management Journal* 41(3):369–399.

Porter ME (1996) What is Strategy? *Harvard Business Review* 74(December):61–78.

Posen HE, Keil T, Kim S, Meissner FD (2018) Renewing research on problemistic search — A review and research agenda. *Academy of Management Annals* 12(1):208–251.

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* 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 (2015) Corporate hierarchy and vertical information flow inside the firm - A behavioral view. *Strategic Management Journal* 36(13):1979–1999.

Relihan T (2018) Agile at scale, explained. *MIT Sloan Ideas Made to Matter* .

Reypens C, Levine SS (2018) Behavior in behavioral strategy: Capturing, measuring, analyzing. Augier M, Fang C, Rindova VP, eds., *Behavioral Strategy in Perspective*, volume 39 of *Advances in Strategic Management*, 221–246 (Bingley, UK: Emerald).

Ries E (2011) *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (New York: Crown Business).

Rietzschel E, Nijstad BA, Stroebe W (2006) Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental Social Psychology* 42(2):244–251.

Rietzschel E, Nijstad BA, Stroebe W (2010) The selection of creative ideas after individual idea generation: Choosing between creativity and impact. *British Journal of Psychology* 101(1):47–68.

Rigby DK, Sutherland J, Noble A (2018) Agile at Scale. *Harvard Business Review* (May–June):88–96.

Rigby DK, Sutherland J, Takeuchi H (2016a) Embracing Agile. *Harvard Business Review* 94(5):40–50.

Rigby DK, Sutherland J, Takeuchi H (2016b) The secret history of Agile innovation. *Harvard Business Review* .

Rijsdijk SA, van den Ende J (2011) Control combinations in new product development projects. *Journal of Product Innovation Management* 28(6):868–880.

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

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

Rosenberg N (1994) Economic Experiments. *Exploring the Black Box: Technology, Economics, and History*, 87–108 (Cambridge, England: Cambridge University Press).

Samuelson L (2004) Modeling knowledge in economic analysis. *Journal of Economic Literature* 42(2):367–403.

Sarasvathy SD (2001) Causation and Effectuation: Toward a Theoretical Shift from Economic Inevitability to Entrepreneurial Contingency. *The Academy of Management Review* 26(2):243–263.

Schilke O, Hu S, Helfat CE (2018) Quo vadis, dynamic capabilities? A content-analytic review of the current state of knowledge and recommendations for future research. *Academy of Management Annals* 12(1):390–439.

Schwaber K, Sutherland J (2013) *The Scrum Guide*.

Scott W (1998) *Organizations: Rational, natural and open systems* (Upper Saddle River, NJ: Prentice Hall), 4th edition.

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

Segars AH, Grover V (1993) Re-examining perceived ease of use and usefulness. *Management Information Systems Quarterly* 17(4):517–525.

Seijts GH, Latham GP, Tasa K, Latham BW (2004) Goal setting and goal orientation: An integration of two different yet related literatures. *Academy of Management Journal* 47(2):227–239.

Selya A, Rose J, Dierker L, Hedeker D, Mermelstein R (2012) A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology* 3:111.

Shalley CE, Perry-Smith JE (2001) Effects of social-psychological factors on creative performance: The role of informational and controlling expected evaluation and modeling experience. *Organizational Behavior and Human Decision Processes* 84(1):1–22.

Shalley CE, Zhou J, Oldham GR (2004) The effects of personal and contextual characteristics on creativity: Where should we go from here? *Journal of Management* 30(6):933–958.

Shaver JM (2005) Testing for mediating variables in management research: Concerns, implications, and alternative strategies. *Journal of Management* 31(3):330–353.

Shelef O, Wuebker R, Barney JB (2020) Heisenberg effects on business ideas. *Available at SSRN 3581255* .

Siggelkow N (2002) Evolution Toward Fit. *Administrative Science Quarterly* 47(1):125–159.

Simon HA (1947) *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (New York: Macmillan).

Simon HA (1955) A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1):99–118.

Simon HA (1962) The Architecture of Complexity. *American Philosophical Society*, volume 106, 467–482.

Simon HA (1964) On the concept of organizational goal. *Administrative Science Quarterly* 9(1):1–22.

Singh J, Fleming L (2010) Lone inventors as sources of breakthroughs: Myth or reality? *Management Science* 56(1):41–56.

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

Spender J, Grant R (1996) Knowledge and the firm: Overview. *Strategic Management Journal* 17:5–9.

Stoop J, Noussair CN, Van Soest D (2012) From the lab to the field: Cooperation among fishermen. *Journal of Political Economy* 120(6):1027–1056.

Strode DE, Huff SL, Hope B, Link S (2012) Coordination in co-located agile software development projects. *Journal of Systems and Software* 85:1222–1238.

Sutherland J, Sutherland J (2014) *Scrum: The Art of Doing Twice the Work in Half the Time* (New York: Crown Business).

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

Taylor A, Greve HR (2006) Superman or the Fantastic Four? Knowledge combination and experience in innovative teams. *Academy of Management Journal* 49(4):723–740.

Thomke S (2003) *Experimentation Matters* (Boston, MA: Harvard Business School Press), 1st edition.

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

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

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

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

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

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

Townsend DM, Hunt RA, McMullen JS, Sarasvathy SD (2018) Uncertainty, knowledge problems, and entrepreneurial action. *Academy of Management Annals* 12(2):659–687.

Tuggle C, Sirmon DG, Reutzel CR, Bierman L (2010) Commanding Board of Director Attention: Investigating How Organizational Performance and CEO Duality Affect Board Members' Attention to Monitoring. *Strategic Management Journal* 968(31):1–43.

Ulrich K (1995) The role of product architecture in the manufacturing firm. *Research Policy* 24(3):419–440.

Ulrich KT, Eppinger SD, Yang MC (2020) (New York: McGraw-Hill), 7th edition.

Unsworth K, Yeo G, Beck J (2014) Multiple goals: A review and derivation of general principles. *Journal of Organizational Behavior* 35(8):1064–1078.

Van de Ven AH, Delbecq AL, Koenig R (1976) Determinants of coordination modes within organizations. *American Sociological Review* 41(2):322–338.

Van den Steen E (2017) A formal theory of strategy. *Management Science* 63(8):2616–2636.

VersionOne (2020) 14th annual state of Agile report.

Vesey JT (1991) The new competitors: They think in terms of 'speed-to-market'. *Academy of Management Perspectives* 5(2):23–33.

Waller MJ, Zellmer-Bruhn ME, Giambatista RC (2002) Watching the clock: Group pacing behavior under dynamic deadlines. *Academy of Management Journal* 45(5):1046–1055.

Walsh JP (1995) Managerial and organizational cognition: Notes from a trip down memory lane. *Organization science* 6(3):280–321.

Wen W, Zhu F (2019) Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal* 40(9):1336–1367.