Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Experimentation and Start-up Performance: Evidence from A/B Testing

Rembrand Koning, Sharique Hasan, Aaron Chatterji

To cite this article:
Rembrand Koning, Sharique Hasan, Aaron Chatterji (2022) Experimentation and Start-up Performance: Evidence from A/B Testing. Management Science

Published online in Articles in Advance 13 Jan 2022

. https://doi.org/10.1287/mnsc.2021.4209

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–20
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Experimentation and Start-up Performance: Evidence from A/B Testing

Rembrand Koning,[a] Sharique Hasan,[b] Aaron Chatterji[b,c]

[a] Harvard Business School, Harvard University, Boston, Massachusetts 02163; [b] Fuqua School of Business, Duke University, Durham, North Carolina 27708; [c] National Bureau of Economic Research, Cambridge, Massachusetts 02138

Contact: rem@hbs.edu, https://orcid.org/0000-0002-6453-0831 (RK); sh424@duke.edu, https://orcid.org/0000-0002-8574-8610 (SH); ronnie@duke.edu, https://orcid.org/0000-0003-1678-9352 (AC)

Received: October 29, 2019
Revised: September 20, 2020; May 20, 2021
Accepted: May 30, 2021
Published Online in Articles in Advance: January 13, 2022

https://doi.org/10.1287/mnsc.2021.4209

Abstract. Recent scholarship argues that experimentation should be the organizing principle for entrepreneurial strategy. Experimentation leads to organizational learning, which drives improvements in firm performance. We investigate this proposition by exploiting the time-varying adoption of A/B testing technology, which has drastically reduced the cost of testing business ideas. Our results provide the first evidence on how digital experimentation affects a large sample of high-technology start-ups using data that tracks their growth, technology use, and products. We find that, although relatively few firms adopt A/B testing, among those that do, performance improves by 30%–100% after a year of use. We then argue that this substantial effect and relatively low adoption rate arises because start-ups do not only test one-off incremental changes, but also use A/B testing as part of a broader strategy of experimentation. Qualitative insights and additional quantitative analyses show that experimentation improves organizational learning, which helps start-ups develop more new products, identify and scale promising ideas, and fail faster when they receive negative signals. These findings inform the literatures on entrepreneurial strategy, organizational learning, and data-driven decision making.

History: Accepted by Toby Stuart, entrepreneurship and innovation.
Funding: R. Koning received funding from the Kauffman Foundation [Grant 201708-2801].
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2021.4209.

Keywords: entrepreneurship • entrepreneurial strategy • experimentation • innovation • organizational learning

1. Introduction

Why do so few start-ups succeed? Scholars often attribute the success and failure of mature companies to differences in strategy: the overarching framework a firm uses to make decisions and allocate resources. In this tradition, credible commitments that force long-term resource allocation decisions provide firms with a competitive advantage (Ghemawat 1991, Ghemawat and Del Sol 1998, Van den Steen 2016). In contrast, recent work suggests that start-ups need a more flexible strategic framework. Levinthal (2017) articulates an organizational learning approach to entrepreneurial strategy with experimentation as the organizing principle. He envisions a "Mendelian" executive who generates alternatives, tests their efficacy, and selects the best course. Similarly, Camuffo et al. (2020) advise entrepreneurs to propose and test many hypotheses about their start-up's strategy to debias learning. Gans et al. (2019) also advocate experimentation but underscore the importance of commitment when choosing between equally viable alternatives.

Although scholars have long appreciated the benefits of an experimental strategy (Bhide 1986, March 1991, Sitkin 1992, Cohen and Levinthal 1994, Sarasvathy 2001, Thomke 2001, Pillai et al. 2020), implementation has historically been costly. Learning from experiments has traditionally been more expensive than other kinds of learning, such as acquiring insights from experience. In recent years, the digitization of the economy and the proliferation of A/B testing tools has led to a decline in the cost of testing ideas (Kohavi et al. 2007, 2009; Kohavi and Longbotham 2017, Azevedo et al. 2020). With this technology, firms of all sizes and vintages can now rapidly implement many experiments to test business decisions and learn from them. Accelerators, venture capital firms, and leading entrepreneurs advocate that start-ups should A/B test nearly everything they do and incorporate what they learn into their strategies.

However, although A/B testing has undoubtedly reduced the cost of evaluating competing ideas, it is an open question whether it facilitates organizational learning and start-up performance. A significant literature suggests that organizations have long learned from a variety of sources, including other firms (Mowery et al. 1996), outside experts (Cohen et al. 2002, Jeppesen and Lakhani 2010), customers (Urban and Von

Hippel 1988, Chatterji and Fabrizio 2014, Dahlander and Piezunka 2014), suppliers (Dyer and Hatch 2004), peers (Chatterji et al. 2019, Hasan and Koning 2019), and even their own failures (Madsen and Desai 2010). The learning process helps firms generate new solutions to their problems, assess their expected benefits, and select the most appropriate course to take. However, prior scholarship also highlights that the learning process is often biased (Denrell and March 2001, Denrell 2003): firms typically generate limited alternatives, rely on ad hoc processes to evaluate ideas, and make decisions based on practicality and instinct rather than data and analysis. Given these challenges to organizational learning, it is not clear whether merely reducing the cost of testing ideas will improve outcomes (Levinthal 2017, Gans et al. 2019, Camuffo et al. 2020). For digital experimentation to matter, firms must also generate many alternatives to test *and* let the results of the A/B tests drive decision making (Brynjolfsson and McElheran 2016).

Even if experimentation matters, how A/B tests impact performance depends on what ideas firms choose to test. There is an ongoing debate among scholars and practitioners about whether A/B tests drive mostly incremental change or enable more significant shifts in product strategy. One characterization of A/B experiments is that they are primarily incremental; for example, a test of a narrow hypothesis about the effect of changing the color of a button or the size of the website's text (Kohavi et al. 2007, Azevedo et al. 2020, Deniz 2021). Although many incremental tests can help firms "hill climb" with an existing product, the process may also blind them to more significant shifts in their industry. Alternatively, it is possible to A/B test significant changes such as the launch of new products, the use of novel recommendation algorithms, or even new business models (Luca and Bazerman 2021). Furthermore, low-cost A/B testing may enable a broader culture of experimentation that leads to significant innovations (Thomke 2020). For example, firms can leverage inexpensive incremental search to reduce the risk of their most significant bets (e.g., a total redesign of a product) by modularizing a significant change into a sequence of smaller testable hypotheses.

We evaluate whether and how experimentation enhances start-up performance. Historically, it has been challenging to address this research question empirically because accurate measurement of economywide experimentation has been prohibitive. We overcome this hurdle by combining four distinct data sources to examine the impact of adopting A/B testing in approximately 35,000 global start-ups over four years. Our data combines a panel of consistent measures of when firms adopt A/B testing with weekly performance measures between 2015 and 2019. We complement this data with a rich set of information about each start-up's technology stack, funding, product characteristics, and website code.

Our first result is that A/B testing improves performance for the start-ups in our sample. Using fixed effects, instrumental variables, and synthetic control models, we find that A/B testing causes an approximately 10% increase in start-up page visits. Our analysis leverages both naturally occurring longitudinal variation in adoption as well as a technological shock—that is, Google's launch of a new A/B testing tool—to identify the effect of A/B testing on performance. This boost in performance appears to compound with time. The effect of testing is roughly 10% in the first few months after adoption and ranges from 30% to 100% after a year of use.

Further, although start-ups of all types in our sample benefit from experimentation, there is heterogeneity in the adoption of A/B testing tools. Just more than 25% of firms with angel or venture capital (VC) investors use A/B testing, whereas only 12.9% of nonfinanced firms do. Silicon Valley start-ups A/B test 25.4% of the time, whereas only 16.1% of non-U.S. start-ups do. Finally, 20.7% of start-ups with more than 10 employees A/B test versus 13% of firms with 10 or fewer workers. Yet we find little evidence that larger VC-backed Silicon Valley firms benefit more from experimentation. Instead, we find suggestive evidence that earlier stage start-ups without financing and those with few employees benefit the most.

We then take an abductive approach and draw on insights from industry practitioners and further quantitative analysis to explore how A/B testing drives organizational learning, product changes, and ultimately significant gains in start-up performance (Timmermans and Tavory 2012, King et al. 2019). Qualitative data from practitioners suggests that A/B testing tools enable not just single experiments, but also more complex experimentation strategies that enable learning about significant product changes and not just incremental improvements. Consistent with our qualitative evidence, we then show that, after adopting A/B testing tools, start-ups are more likely to change their website's code, make both incremental and significant code changes, and introduce new products. Indeed, firms that adopt A/B testing introduce new products at a 9%–18% higher rate than those that do not experiment. Finally, consistent with the idea that A/B testing improves organizational learning, we find evidence that A/B testing is associated with an increased likelihood of tail outcomes—that is, more zero page-view weeks and more 50k+ page-view weeks.

These findings advance our understanding of entrepreneurial strategy, organizational learning, and the impact of digitization on business. We empirically demonstrate that an experimental approach to strategy

leads to significant performance improvements for the start-ups we analyze (Gans et al. 2019, Camuffo et al. 2020). Our work suggests that we should reconsider the classic debate between emergent and intentional approaches to strategy in the context of entrepreneurship (Porter 1980, Mintzberg 1990). In our sample, we find that experimentation helps drive both valuable incremental changes *and* the development of significant product improvements. The Mendelian executives envisaged by Levinthal (2017) can use A/B testing to find a "middle ground" between the two dominant views of strategy, allowing their firms to reap the benefits of both approaches. This insight places organizational learning via experimentation at the heart of entrepreneurial strategy and suggests that we should reconsider the dominant characterization of A/B testing as being solely tactical.

We also contribute to a growing literature on the digitization of the economy. As the analytical capabilities of large firms increase, so does their productivity (Brynjolfsson and McElheran 2016, 2019). In the spirit of recent studies that endorse an entrepreneurial process rooted in the scientific method (Camuffo et al. 2020), we demonstrate that start-ups also benefit from data-driven decision making. Finally, our results echo a related marketing literature highlighting the importance of market testing for improving products (Urban and Katz 1983, Boulding et al. 1994, Runge and Nair 2021).

2. Theoretical Framework

2.1. Experimentation as an Entrepreneurial Strategy

Uncertainty is endemic to the entrepreneurial process (McMullen and Shepherd 2006). Entrepreneurs must make many decisions, often with risky or unknown payoffs (e.g., McDonald and Eisenhardt 2020, Ott and Eisenhardt 2020). They must choose which customers to serve, what product features to include, and which channels to sell through (McGrath and MacMillan 2000). What framework should an entrepreneur use to make these decisions?

Recent research in strategic management theorizes that organizational learning via experimentation is a promising approach for entrepreneurial strategy (Levinthal 2017, Gans et al. 2019, Camuffo et al. 2020). In this work, experimentation is cast as a three-part process: Entrepreneurs first *generate* ideas to introduce variation in the number and nature of strategic options. Next, they *test* the viability of selected options. Finally, they must *make decisions* based on the test results. An experimentation framework biases entrepreneurs toward learning and adaptation, avoiding premature or costly commitments (Bhidé 1986, 2003).

Although experimentation has long been promoted as a framework for strategic decision making by academics (Thomke 2001, Bhidé 2003) and practitioners (Ries 2011, Blank 2013, Kohavi and Longbotham 2017), it has traditionally been costly to implement (March 1991). Generating new ideas is difficult and potentially diverts effort and resources away from other essential tasks. Even more challenging than creating many new ideas is evaluating them all (Knudsen and Levinthal 2007). Running rigorous experiments on new product features, for example, requires a flexible production process, requisite scale to test various options, and the capability to interpret the results. Further, deciding between viable options is also challenging (Simon 1959), and bureaucracy and other sources of inertia inside organizations may hinder the ability to take decisive action (Hannan 1984). Finally, there are many, arguably less expensive, alternative channels from which firms can learn. As mentioned, firms have traditionally learned from their own experience, competitors, or other readily available sources, making investments in formal experimentation less attractive. Given the factors described, firms have rarely used formal experiments to inform business decisions.

However, in the last decade, rapid digitization of the global economy has altered this calculus (Brynjolfsson and McAfee 2012). In particular, the cost of running controlled tests that compare alternatives has declined dramatically (Kohavi et al. 2007, Kohavi and Longbotham 2017). One key driver of this transformation is that experimenting with product features on a website, whether on an e-commerce or enterprise platform, is much less costly than in a manufacturing process. Furthermore, the scale afforded by digital businesses allows these companies to run many simultaneous and independent tests. Finally, advances in data analytics enable firms to interpret the results of their experiments reliably (Brynjolfsson and McElheran 2016). Collectively, these tests have come to be known as A/B tests (Azevedo et al. 2020). Today, software products such as Optimizely and Google Optimize allow any firm with a digital presence to set up controlled experiments and analyze the data using prepackaged software.

Although no prior published studies examine the benefits of A/B testing across many firms, various scholarly and practitioner accounts have described the utility of experimentation inside organizations (Kohavi et al. 2007, 2009; Xu et al. 2015; Kohavi and Longbotham 2017). Online retailers, for example, test different bundling, pricing, and product display strategies (Sahni et al. 2016, Dubé et al. 2017). Networking platforms experiment with social features, recommendation algorithms, and content to increase user engagement (Aral

and Walker 2011, 2014; Kumar and Tan 2015; Bapna et al. 2016). Media companies A/B test the placement of articles or videos on their website, title variants, and subscription prices (Gomez-Uribe and Hunt 2016, Lawrence et al. 2018).

2.2. A/B Testing, Learning, and Performance

Nevertheless, organizational learning requires *more* than just a reduction in the cost of testing ideas (Fabijan et al. 2017), and this is the crucial innovation facilitated by Optimizely and other A/B testing platforms (see, for example, Siroker et al. 2014). Recall that learning via experimentation has three parts: the introduction of variation, the testing of alternatives, and selecting candidate solutions (Levinthal 2017). If A/B testing directly reduces the cost of testing ideas, how might we expect it to influence organizational learning more broadly?

Prior research suggests that when the cost of a vital input declines, organizations often respond by investing in complementary practices (Nordhaus 2007). For example, researchers have documented that investments in information technology yielded returns for firms only when they invested in hiring workers with relevant expertise (Brynjolfsson and Hitt 2003). Likewise, the reduced cost of testing ideas may incentivize a firm to increase idea generation. John Cline, director of engineering at Blue Apron, highlights how an A/B testing platform led to more product ideas[1]: "Now that we have this capability, other groups have started using it. We went from one or two teams doing one or two tests a quarter to now, when we probably have at least 10 tests live at any given moment and a large number of tests every quarter being run by every product team."

Another way that A/B testing supports idea generation is by reducing the impact of failed ideas and improving execution. For example, Emily Dresner, the CTO of Upside Travel, notes,[2] "We can ship MVPs and eliminate poor paths—bad messages, bad landing pages, bad flows—without jeopardizing our current progress."

We expect A/B testing to facilitate organizational learning through a variety of channels. Directly, a firm learns about the quality of any idea it tests. Indirectly, A/B testing improves learning by increasing incentives for idea generation and execution (Levinthal 2017, Gans et al. 2019).

Organizational learning is essential for firms because it has long been linked to competitive advantage and better performance (March 1991). The firms that learn fastest are more likely to build and sustain an edge on their competitors, allowing them to solve complex business challenges and develop new products more quickly. Yet the impact of learning on performance depends on the kinds of experiments a firm runs. Testing only incremental changes may yield fewer insights than conducting significant experiments.

2.3. The Alternative to Formal Experimentation

Before proceeding to our empirical approach, we consider the appropriate counterfactuals for A/B testing. If start-ups are not conducting formal experimentation, what other strategies are they undertaking for organizational learning? We briefly review two approaches that have been highlighted in prior work. First, extensive literature in entrepreneurship documents that founders are overconfident in assessing the quality of their ideas (Camerer and Lovallo 1999) and are vulnerable to confirmation bias in decision making (Nickerson 1998, McGrath 1999). The implication is that such entrepreneurs invest time and effort into implementing strategies that will likely fail (Camuffo et al. 2020). This approach is less effective than experimentation, ultimately leading to significant performance differences between firms that experiment and those that do not.

Next, prior work documents that firms have long learned from "uncontrolled experiments" or tweaks to their product development process (David 1975). Hendel and Spiegel (2014) attribute much of the substantial productivity gains in a steel mill they studied over 12 years to learning from uncontrolled experiments. These tweaks include experimenting with how scrap enters the furnace and the timing of various production tasks. Levitt et al. (2013) document a similar phenomenon in an automaker's assembly plant in which learning by doing led to productivity gains (Arrow 1962). In our sample of high-technology start-ups, firms that are not conducting formal experiments may be tweaking their products informally, which may well lead to learning and improved performance, albeit at a slower pace. A/B testing should reduce the false positives of confirmatory search and accelerate the rate of discovering product improvements compared with tweaking. If, however, tweaking does lead to sustained gains in performance, A/B testing might have only a muted effect on firm performance.

In the next section, we evaluate whether A/B testing leads to improved performance for a large sample of high-technology start-ups. After establishing this relationship, we explore how experimentation via A/B testing can yield performance gains. Specifically, we use qualitative and quantitative evidence to document whether A/B testing is used to test incremental or significant changes to products.

3. Data and Methods

To test whether A/B testing improves start-up performance, we construct a longitudinal data set comprising 35,262 high-technology start-ups founded between 2008 and 2013. Our data include information about these start-ups compiled from four distinct sources. Crunchbase provides us detailed information

about each start-up's product, funding status, age, location, and team size. We complement the Crunchbase information with weekly measures of page views/visits for each start-up from SimilarWeb and use BuiltWith to gather information on the technologies the start-ups use to build their product, most notably whether and when they use A/B testing tools. Finally, for just under a quarter of our start-ups, we can collect data on their homepage code over time from the Internet Archive's Wayback Machine to measure the degree and nature of change associated with adopting A/B testing. Online Appendix A1 describes each data source in detail.

We link start-ups across these data sources through website URLs. Unlike firm names, URLs are unique identifiers, eliminating the need for fuzzy matches.[3] To ensure that our sample begins with "active" start-ups likely to adopt A/B testing tools, we include only start-ups in the Crunchbase data with nonzero page views in March 2015, the first month for which we have SimilarWeb data. We also exclude start-ups that have subdomains—versus primary domains—as URLs because SimilarWeb does not provide independent estimates for subdomains.[4] Finally, some start-ups consist of thousands of subdomains. In BuiltWith, for example, technologies used by subdomains are attributed to the parent domain (e.g., wordpress.com would be assigned any technology associated with my-awesome-blog.wordpress.com). To address this problem, we exclude pages with more than 80 active unique technologies as of March 2015.

After these exclusions, our primary data set consists of 35,262 independent product-oriented start-ups founded between 2008 and 2013. Our panel captures the characteristics, web metrics, and technology adoption trajectories of these start-ups starting in the week of April 5, 2015, until March 24, 2019—amounting to 208 weeks (four years)—and a total of 7,334,496 firm-week observations.

We organize our analysis of this rich data set into two distinct parts. First, we study whether A/B testing impacts start-up performance through a series of sequentially developed models. After presenting these results, we use abductive reasoning and a second set of analyses to explore how A/B testing impacts start-up performance. In this later part of our analysis, we iterate between theorizing, data analysis, and data collection to arrive at the most likely explanation for why we observe that A/B testing improves start-up performance.

4. Does A/B Testing Impact Start-up Performance?

Our first set of analyses estimates the impact of A/B testing on start-up performance for the 35,262 in our sample. First, we describe our primary variables. We then present results from standard two-way fixed effect (TWFE) and event study models. Our large sample and long panel allow us to estimate various models to check robustness, explore how the impact of A/B testing varies over time, and investigate heterogeneity in when A/B testing matters most. We then use the March 2017 launch of a new A/B testing tool, Google's Optimize and Optimize 360, as a shock that allows us to better identify the causal effect of A/B testing on start-up growth. We leverage this shock to estimate models using instrumental variables, TWFE, and synthetic control approaches. We find robust evidence that A/B testing significantly improves start-up performance and that this performance effect compounds with time.

4.1. Variable Construction

Using A/B tool, our primary independent variable, is constructed from the BuiltWith technology data by identifying the set of tools that focus on website A/B testing. Our final set of A/B testing technologies includes the following tools: AB Tasty, Adobe Target Standard, Experiment.ly, Google Optimize, Google Website Optimizer, Omniture Adobe Test and Target, Optimizely, Optimost, Split Optimizer, and Visual Website Optimizer.[5]

Table 1 shows that just under 18% of firms use an A/B testing tool in the 208 weeks of our data. On average, 8% of firms actively use A/B testing technology in any given week. In our data, Optimizely is the market leader, accounting for just more than 60% of the weeks in which firms are A/B testing. The next most prominent A/B testing software is Google with slightly more than 20% of the market. The remaining 20% is split between Visual Website Optimizer, Adobe, AB Tasty, and Experiment.ly.

Table 1 also reveals significant heterogeneity in the types of firms that use A/B testing tools. Just more than 25% of firms with angel or VC investors use A/B testing, whereas only 12.9% of nonfinanced firms do. Silicon Valley start-ups have an adoption rate of 25.4% compared with only 16.1% for non-U.S. start-ups. Further, 20.7% of start-ups with more than 10 employees have adopted A/B testing versus 13% for firms with 10 or fewer workers. Overall, it appears that start-ups that are larger and have financial backing are more likely to adopt experimentation tools.

Technology Stack measures technology adoption in addition to A/B testing software. For each week, we calculate the number of distinct non-A/B testing tools active on the website, according to BuiltWith, at the start of the week. Over the 208 weeks, some firms drop to five technologies (5th percentile), and others grow in complexity, reaching 111 different web technologies (99th percentile). To account for the skewness in the technology

Table 1. Our Panel Covers 35,262 Start-ups for 208 Weeks (Four Years)

Panel A: Start-up–week level

	Mean	Median	Standard deviation	Minimum	Maximum	*N*
Using A/B tool?	0.08	0.00	0.27	0	1	7,334,496
Log(Visits + 1)	6.11	6.64	3.73	0	20	7,334,496
Log(Technology Stack + 1)	3.49	3.69	0.86	0	5.81	7,334,496

Panel B: Start-up level

	Number of start-ups	Percentage A/B testing
Not angel/VC funded	22,250	12.9
Angel/VC funded	13,012	25.2
Founded 2012–13	14,569	15.3
Founded 2010–11	11,966	16.5
Founded 2008–09	8,727	15.2
Outside United States	14,645	16.1
In United States, outside Bay Area	12,493	18.9
Bay Area	4,187	25.4
1-10 employees	15,393	13.0
11+ employees	19,840	20.7
Fewer than 1,500 weekly visits	17,189	8.1
More than 1,500 weekly visits	18,073	26.3
Commerce and shopping	4,517	24.1
Advertising	2,445	14.8
Internet services	2,079	17.2
Software	2,047	16.1
Data and analytics	1,940	21.6
Apps	1,746	17.1
Content and publishing	1,579	14.8
Financial services	1,547	23.6
Education	1,386	19.3
Information technology	1,233	20.0
Healthcare	1,042	19.2
Hardware	1,030	16.5
Other	12,671	14.2

Notes. Panel A provides summary statistics at the startup-week level. Panel B shows the number of startups of each type and the percent that use an A/B testing tool for at least one week during our panel.

adoption data, we log this variable. However, results are unchanged when we include the raw counts.

Log(Visits+1) is the log of the weekly page visits as estimated by SimilarWeb. Because page views can drop to zero, we add one before transforming the variable.

4.2. Two-Way Fixed Effects Performance Effects

We begin by assessing the impact of A/B testing on growth by estimating a TWFE model:

$$Y_{it} = \beta(A/B\,Testing_{it}) + \theta(Technology\,Stack_{it}) + \alpha_i + \gamma_t + \epsilon_{it}, \quad (1)$$

where Y_{it} is logged visits, and our interest is in β, the impact of A/B testing adoption on start-up performance. To reduce selection bias, the model includes fixed effects for each week (γ_t) to control for observed and unobserved nonparametric time trends. Such trends could consist of changes to general economic conditions and an increase in internet usage or access as well as a host of other time-varying factors that could bias our estimates. Second, we include firm fixed effects α_i to control for time-invariant differences between firms. These factors could consist of the quality of the initial start-up idea, the existence of a specific strategy, location advantages, and founders' educational backgrounds, among other fixed resources or capabilities of the start-ups.

In addition to our fixed effects, we include a weekly time-varying control variable for the number of other technologies adopted by the start-up. Including this time-varying control increases our confidence that observed differences in performance attributed to A/B testing are not derived from other changes to a company's technology stack (e.g., adding Facebook's Tracking Pixel).

Table 2, Model 1, estimates the raw correlation between A/B testing and weekly visits after accounting for week fixed effects. We find that firms that use A/B testing have 296% more visits than those that do not. In Model 2, we include technology stack controls. Suppose the adoption of A/B testing is correlated

with the adoption and use of other technologies (e.g., payment processing tools). In that case, the raw estimate might reflect the impact of different technologies and not of A/B testing or major technological pivots that lead to better performance. The estimate drops to 19%, but the result remains precisely estimated and significant at the 0.1% level. Controlling for time-varying technology adoption captures a meaningful amount of firm heterogeneity. In Model 3, we account for firm-level heterogeneity by including firm fixed effects. The estimated impact of A/B testing drops to 55%. Finally, Model 4 includes both our technology stack control and firm fixed effects. The estimate remains statistically significant with a magnitude of 13%. The point estimate and standard errors suggest that A/B testing improves start-up performance for the firms in our data.

4.3. Alternative Specifications and Robustness Checks

To further assess the robustness of our results, in Figure 1, we present estimates of the effect of A/B testing using a variety of different modeling choices.[6] The leftmost estimate (the circle) presents the estimate from Table 2, Model 4. Moving right, the next estimate (the diamond) shows the estimates swapping out our logged-plus-one dependent variable for the inverse hyperbolic sine transformation. The choice of transformation does not alter the estimate. The third estimate (the square) excludes all observations in which the number of visits that week is zero. In our balanced panel, if a firm fails quickly—and so its visits go to zero quickly—it never adopts A/B testing tools, but all postfailure observations are still included. By excluding zeros, we end up with an unbalanced panel in which observations are censored if they fail. The estimate remains unchanged. It does not appear that the overrepresentation of zero observations drives our findings.

The next three coefficients repeat the same pattern and include an additional firm-week slope fixed effect for each start-up. Including slope fixed effects allows us to address the possibility that our finding is a consequence of fast-growing start-ups adopting A/B testing at higher rates. Although the estimate shrinks to between 5% and 10%, it remains significant at the 1% level.

The estimates thus far control for variation in growth rates, but they mask time-varying differences in the impact of A/B testing on growth. For example, it could be that A/B testing has an immediate and small effect on growth, or perhaps its effect is gradual and compounds with time from adoption. Our TWFE results represent the average difference (within firms) between observations when A/B testing is used and firm-weeks when A/B testing is not. In essence, the TWFE can be thought of as an average of a multitude of two-period (e.g., week 1 versus 26, week 1 versus 52, etc.) by two-group (i.e., using versus not using A/B testing) comparisons (Goodman-Bacon 2021).

To test if the impact of A/B testing increases over time, in the column "TWFE 2×2" in Figure 1, we show three estimates from models that include two periods of our data. The first (the triangle) comes from a simple difference-in-differences model using only data from the 1st and 26th weeks of our panel. As with all our models, it includes the technology stack control and firm and week fixed effects. The estimate is noisy, but the magnitude is similar to our baseline estimate at about 10%–15%. The second estimate (the x) compares the 1st week to the 52nd week in our panel. The point estimate is more substantial at about 20%. The third estimate focuses on our first and last weeks of data and suggests that A/B testing results in 30% more weekly visits over a four-year horizon. It appears that the impact of A/B testing increases with time.

The last nine estimates in Figure 1 replicate the first nine but include only start-ups that transitioned to or

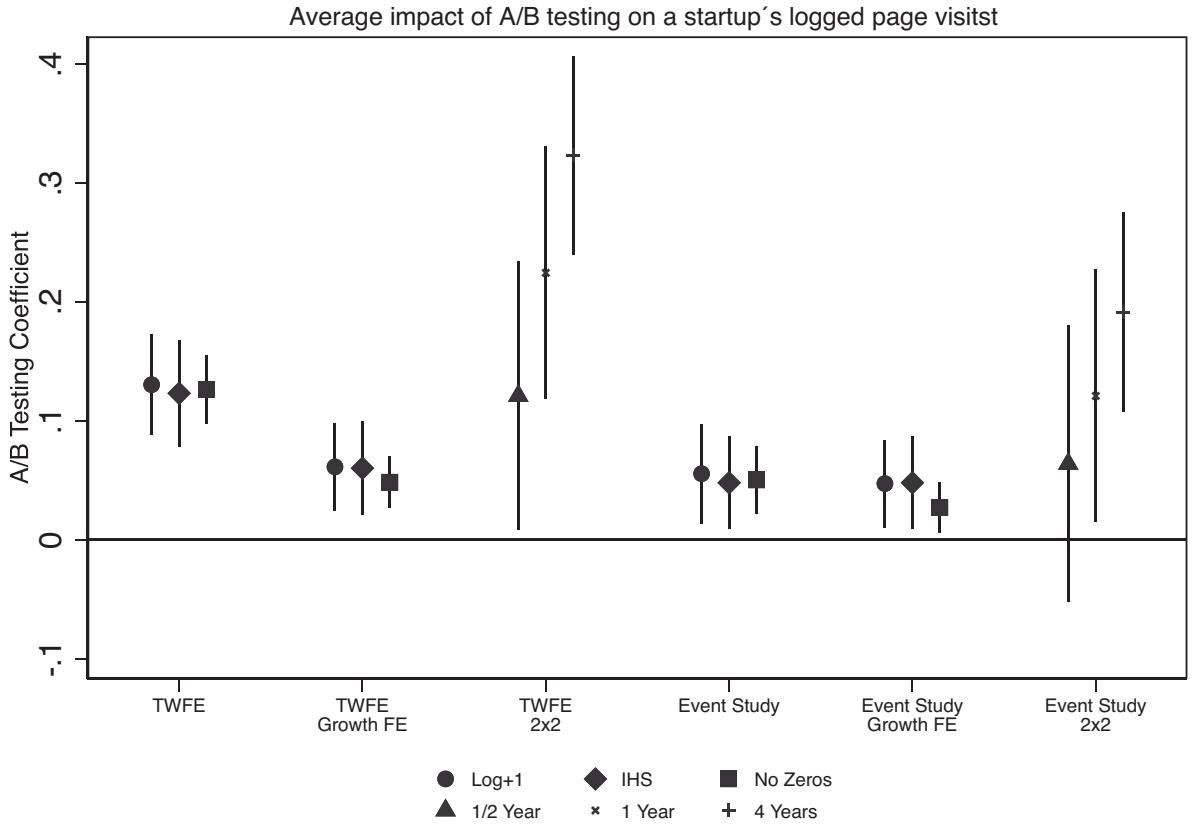
Table 2. Panel Regressions Showing How the Inclusion of Firm Fixed Effects and Time-Varying Controls for a Firm's Technology Stack Reduce the Estimated A/B Testing Effect from Nearly 300% to Just more than 13%

	(1)	(2)	(3)	(4)
	Log(Visits + 1)			
Using A/B tool?	2.957***	0.190***	0.553***	0.131***
	(0.046)	(0.024)	(0.026)	(0.022)
	[2.866, 3.047]	[0.143, 0.237]	[0.503, 0.604]	[0.088, 0.173]
Observations	7,334,496	7,334,496	7,334,496	7,334,496
Number of firms	35,262	35,262	35,262	35,262
Number of eeeks	208	208	208	208
Week fixed effects	Y	Y	Y	Y
Firm fixed effects			Y	Y
Technology stack control		Y		Y

Notes. Weekly data from 2015 to 2019 on 35,262 Crunchbase start-ups founded between 2008 and 2013. Linear regressions with robust standard errors clustered at the firm level in parentheses. Brackets show 95% confidence intervals.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Figure 1. The Effect of A/B Testing on Log Page Visits Holds Across a Range of Model Specifications



Notes. "TWFE" indicates standard two-way fixed effects models, "Growth FE" indicates the model includes firm growth fixed effects, and "2 × 2" indicates that the estimate is from is a simplified difference-in-differences model that includes only data from the first week in our panel and a single observation either a half-year, year, or four years later. "Event study" indicates that only A/B switchers are included in the data. "IHS" indicates that we use the inverse hyperbolic sine instead of logged-plus-one visits. "No Zeros" indicates that all weeks in which page views are zero have been excluded from the data. All models include start-up fixed effects, week fixed effects, and a control for the size of the start-up's technology stack. Bars are 95% confidence intervals.

from A/B testing.[7] These "event study" models rely exclusively on comparisons between firms that will adopt or have adopted. The estimates are somewhat smaller but still greater than zero and also suggest that the impact of A/B testing increases with time.

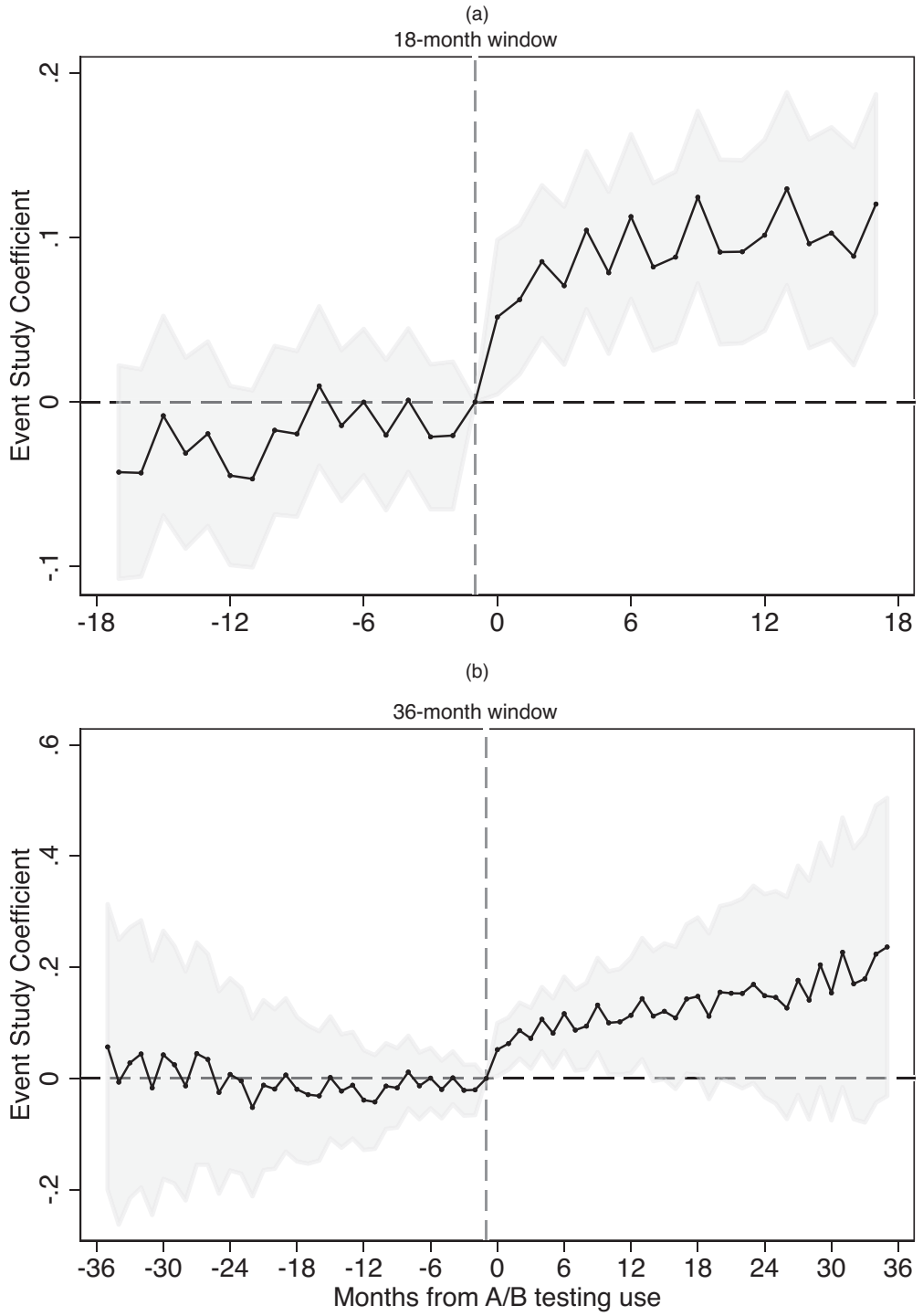
Using the event study specification also allows us to generate lead–lag plots showing the estimated effect size relative to when a firm starts and stops using A/B testing. To build these lead–lag plots, we first aggregate to the monthly level so that it is easier to see if an estimate and trends are shifting in statistical significance. Using these monthly data, we then include dummy variables for the number of months before and after the firm switches to using or not using A/B testing tools. As with our TWFE estimates, our event studies are identified using firms that either adopt A/B testing once in our panel or stop using A/B testing at one point in our panel.

The top panel in Figure 2 provides estimates from a model that includes dummies for the 18 months before and after the switching event. The month before the event serves as the excluded baseline. Before a firm switches, there are few pretrends, and the estimates overlap with zero. There are many reasons why we might expect a lack of observable pretrends. If sales cycles drive adoption, it might lead firms to adopt A/B testing tools at points unrelated to other firm events. If an A/B testing provider significantly raises its tool's price, which firms such as Optimizely have done in the past, firms may choose to stop using the tool for reasons unrelated to any firm-specific shocks. Although we can never rule out unobserved differences in pretrends with nonexperimental data, the lack of any trends before the event date suggests adoption and disadoption decisions may be plausibly exogenous.

The effect of A/B testing appears to grow with time although the estimates in our event study graph are noisy. That said, they are consistent with our 2 × 2 models, which more strongly suggests that the value of A/B testing increases over time.

The bottom panel in Figure 2 shows estimates for a model that includes dummies for 36 months before

Figure 2. Event Study Plot Showing the Effect of A/B Testing over Time



Notes. Panel (a) shows the effect 18 months before and after use, and Panel (b) shows a 36-month window before and after. The month before the event serves as the excluded baseline. The *y*-axis is the estimated coefficient for the A/B testing effect. All models include the technology stack control and account for week and firm fixed effects. The shaded region indicates 95% confidence intervals.

and after the use of A/B testing. This extended event study again reveals little in the way of pretrends. Although the estimates are noisy, the value of A/B testing increases with time. After three years, the impact of A/B testing is close to 30% although the 95% confidence intervals are wide.

We report several additional robustness checks in the online appendix. In Online Appendix A3, we

extend the robustness checks in Figure 1 by showing that changes to our definition of what constitutes an A/B tool do not alter our findings. When we use a more expansive definition of A/B testing software that includes tools that allow for A/B testing but whose focus is on general analytics, we still find a very similar pattern of results. Online Appendix A4 presents placebo estimates using the adoption of cloud font tools to check if our preferred modeling strategy in Table 1 (column (4)) mechanically leads to positive effects. We find that, once we include our technology stack control and firm fixed effects, the estimated impact of cloud font tools is a precisely estimated zero. In Online Appendix A5, we show that our pattern of findings holds when we model our data as a dynamic panel instead of using the "static" difference-in-differences setup shown in Table 1. These dynamic models suggest that A/B testing improves long-run growth by anywhere from 10% to 200%.

Finally, in Online Appendix A6, we test within our sample for heterogeneous performance effects of A/B testing. We find that the positive effects of experimentation are relatively stable for start-ups in different verticals (e.g., financial services, education, hardware, e-commerce) and at different stages. However, it does appear that smaller, younger, and non-VC-backed start-ups see more meaningful gains than larger, older, and VC-backed firms. Intriguingly, when compared with the adoption results in Table 1, it appears that the firms that are most likely to benefit are the least likely to adopt. This finding suggests that the lower adoption rates for younger firms without financial backing are not a result of the lack of potential benefits but are rooted in frictions in implementation.

4.4. Using the Launch of Google Optimize to Estimate Instrumental Variable and Synthetic Control Models

Although the evidence thus far points to A/B testing improving start-up growth, it is not without limitations. First, we do not have a sense of why firms choose to use (or not use) A/B testing tools. Although our models account for many forms of selection bias, the specter of unobserved time-varying shocks remains. Second, our TWFE and event study estimates pool together many different tools and decisions to adopt and disadopt. Although this approach increases our statistical power and the generality of our results, causal inference remains a challenge because of the nonrandom adoption of A/B testing technologies.

To address these concerns, we focus on the global launch of Google's Optimize and Optimize 360 A/B testing suite on March 30, 2017. The tool, which integrates on top of Google Analytics, is similar to competitors such as Optimizely and AB Tasty. Beyond providing statistical estimates, the tool's interface makes it possible for product managers and marketers to deploy experiments with less engineering support. The Google Optimize suite offers both free and paid tiers. We use the launch as a shock that allows us to estimate both instrumental variable and synthetic control estimates for the impact of A/B testing on start-up performance.

We use this shock in two ways: both across and within firms. First, to construct an instrument for Google Optimize adoption, we use the fact that firms using Google's Tag Manager (GTM) software adopt Google Optimize at much higher rates. Google Optimize strongly recommends—although it doesn't require—firms to use GTM software to get the most out of Google Optimize. Thus, start-ups that already use GTM should be more likely to adopt Google Optimize because they only need to install and learn one new tool instead of two new technologies. Indeed, we find that firms that had adopted GTM as of March 2016, a year before the launch of Google Optimize, adopt at twice the rate of non-GTM users. This stylized fact, combined with fixed effects, creates an effective difference-in-differences instrumental variable identification strategy. We use whether a start-up had adopted GTM a year before the launch as an instrument that increases the likelihood that a firm will use Google Optimize while controlling for firm and time fixed effects. As discussed in detail in Online Appendix A7, we find that adopting Google Optimize consistently improves performance by well over 100%. Similarly, when we use our simple TWFE model, we find that adopting Google Optimize improves the average number of visits by more than 30% (Online Appendix A8).

To further test the robustness of these larger estimates, we also combine the Google Optimize shock with synthetic control models to estimate the causal impact of A/B testing. The shock occurs two years into our panel, which allows us to leverage at least 100 weeks of preadoption data to construct matched synthetic controls for all Google Optimize adopters. Crucially, because no start-up can adopt Google Optimize during this period, we can be confident that early Google Optimize adopters have not already selected out of our sample, nor have they adopted before we had enough observations to build a synthetic control. We use these synthetic controls to trace the expected growth trajectory had these start-ups not adopted the Google Optimize A/B testing tool. An additional benefit of this approach is that, because we construct time-varying counterfactuals, we can improve on our event study and 2×2 estimates to directly test if the effect of A/B testing appears to increase with time since adoption.

Specifically, we use generalized synthetic controls (Xu 2017). This method has the advantage that it

naturally averages across many different treated units, whereas the canonical synthetic control of Abadie et al. (2010) builds a counterfactual for only one treated unit. We use cross-validation to select the optimal number of factors for constructing the synthetic counterfactual for this method. To estimate uncertainty intervals, we use a nonparametric bootstrap procedure with 500 runs. Finally, we include our time-varying technology stack control and firm and week fixed effects as with all our models.

Figure 3 shows the estimated treatment effect relative to the time of Google Optimize adoption. Panel A focuses on the effect trajectory a year before and after adoption; Panel B shows the complete set of estimates that run two years before and after. The fact that the estimates are near zero before adoption suggests that the model generates precise preadoption counterfactuals.[8]

Turning to the postadoption period, we see a small but noisy effect initially. Unlike our event study estimates, there is no clear and immediate increase. With time, the impact grows, and by six months out, the effect is significant. The plot suggests that the impact of A/B testing increases with time since adoption although Panel B indicates that it stabilizes a little after one year.

Table 3 provides point estimates from the model. In column (1), we report the estimated effect of the week before adoption. The estimate is small and near zero (−0.1%), consistent with the idea that the model adequately captures pretrends. By 26 weeks (column (2)), the estimate is 37% and significant at the 5% level although the 95% confidence intervals range from 5.8% to 64.8%. By week 52, the estimate is 128% and again significant. Further, it is significantly higher than the estimate at six months. Finally, column (4) reports the average treatment effect over the postadoption period. The estimate is 67.6% with 95% confidence intervals ranging from 24.6% to 93%.

Beyond providing further robustness checks, the synthetic control analysis results help explain the variation in effect sizes we see with the 2×2, TWFE, event study, and IV analyses. If the impact of A/B testing grows with time, as it appears to, then comparing the effect over longer time horizons leads to more long-term testers that experience more substantial treatment effects. The TWFE models rely on a large panel that includes firms that adopt A/B testing for long periods and firms that install and quickly uninstall the technology. If the impact of A/B testing takes time to materialize and the TWFE includes more short-term adopters, this would lead to smaller estimates of the treatment effect. Suppose the IV compliers are more likely to adopt the tool for the long term, for example. In that case, because it integrates relatively seamlessly with the GTM tool on which they already rely, we should expect the IV estimates to be larger. Overall, our findings suggest that researchers need to be aware that treatment effects may take time to materialize when analyzing the impact of experimentation practices.

In summary, these findings suggest that, for the firms in our data, A/B testing improves start-up performance and increasingly so with time. Moreover, A/B testing appears to enhance the performance of a wide array of start-ups, including those inside and outside of entrepreneurial hubs and in many industries, ranging from e-commerce to education.

5. How Does A/B Testing Impact Start-up Performance?

Although our first set of results lends credence to the idea that A/B testing substantially increases start-up performance, these results provide little insight into the changes firms make to achieve these improvements. If A/B tests are incremental, how could the adoption of A/B testing lead to such significant performance gains? Although we cannot directly observe the A/B tests firms are running, we provide additional qualitative testimony from start-ups and quantitative evidence on code and product changes to better understand how this tool could significantly impact firm strategy and performance.

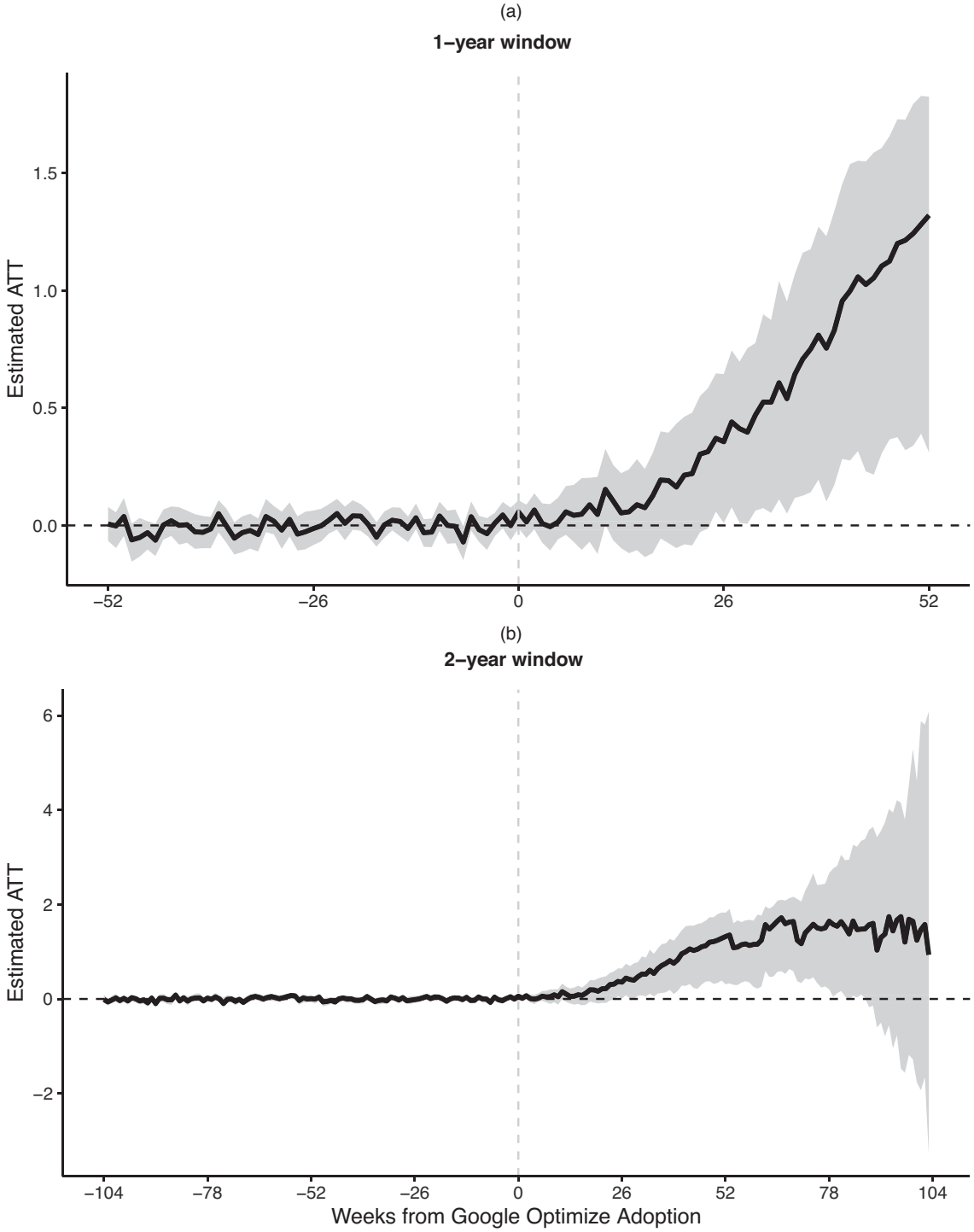
5.1. Can A/B Tests Lead to Significant Product Changes?

Practitioners suggest that A/B testing is used for more than incremental experimentation (e.g., button colors or landing page copy). Instead, managers and engineers routinely note how experimentation has become a core part of their product development and deployment strategies. For example, engineers at Auth0 recommend an A/B testing strategy in which start-ups "go big, then go small," lest they get stuck in a local maximum[9]: "When looking to A/B test, it's better to make big changes first to have the most significant effect. Small changes are good for optimizing further down the road, but those first need to be guided by larger-scale tests. In any industry, starting high-level and testing the entire landing page can open many doors for things to A/B test further down the line. Making big changes can reveal surprising and maybe even unintuitive results that you might've never expected."

This reflection suggests that A/B testing helps break down larger business challenges into smaller, testable hypotheses. By sequentially testing the most critical ideas, product managers can progress toward long-term objectives with uncertain payoffs.

In addition, A/B testing may make the costs and benefits of innovation more transparent, helping managers build a robust quantitative case for a strategic shift. Joel Lewenstein, a product designer at the data collaboration software start-up Airtable, describes the

Figure 3. The Estimated Average Treatment Effect on Treated for the Adoption of Google's Optimize A/B Testing Tool on Logged Page Visits



Notes. Estimates are from a generalized synthetic control model. Panel (a) shows the estimated difference between an adopter and a synthetic control a year before and after the firm adopts. Panel (b) shows the same effects but over a longer two-year prewindow and postwindow. The shaded region shows 95% confidence intervals.

value of quantification through A/B testing in helping manage the trade-offs that are a consequence of making big changes[10]: "Even the best qualitative theory rarely includes predictions about degree of impact. Testing and confirming improvements results in an outcome that can be weighed against other potential changes and balanced against the work necessary to build and support that change."

Table 3. Point Estimates from a Generalized Synthetic Control Model Show That Start-ups That Adopt the Google's Optimize A/B Testing Tool See More Growth and That This Growth Increases with Time

	(1)	(2)	(3)	(4)
		Log(Visits + 1)		
	1 week before	26 weeks after	52 weeks after	Average
Using Google A/B tool?	−0.001	0.371*	1.282*	0.676*
	(0.037)	(0.159)	(0.378)	(0.191)
	[−0.063, 0.076]	[0.058, 0.648]	[0.391, 1.829]	[0.246, 0.930]
Observations	4,359,264	4,359,264	4,359,264	4,359,264
Adopting firms	618	618	618	618
Week fixed effects	Y	Y	Y	Y
Firm fixed effects	Y	Y	Y	Y
Technology stack control	Y	Y	Y	Y
Number of factors	5	5	5	5

Notes. Weekly data from 2015 to 2019 on 20,958 Crunchbase start-ups that have yet to adopt A/B testing at the time of the launch of Google 360 in March 2017. Estimates are from a single generalized synthetic control model in which the number of unobserved factors was selected by a cross-validation procedure. Standards errors and confidence intervals calculated using a nonparametric bootstrap with $N = 500$. Standard errors in parentheses. Brackets show 95% confidence intervals.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Further, aside from quantitative data, the existence of A/B testing tools provides further confidence that firms can learn in new or uncertain environments. Consider a company contemplating a market-entry decision. If company executives are confident that they can quickly learn about customers in their new market and adapt using A/B testing, they may be more likely to enter in the first place. By accelerating learning after big strategic decisions, A/B testing can enable significant changes in the first place. Xu (2015) notes that quantification through A/B helps derisk strategic bets: "For example, when we make a strategic bet to bring about a drastic, abrupt change, we test to map out where we'll land. So, even if the abrupt change takes us to a lower point initially, we are confident that we can hill climb from there and reach a greater height through experimentation."

Moreover, some practitioners suggest that the real benefit of experimentation through A/B testing is the ability to conduct clearly defined experiments on major changes to a product. EJ Lawless, a cofounder of Experiment Engine,[11] reflects on thousands of A/B tests run on the company's platform[12]: "We also wanted to explore whether incremental changes (like Google's test of 41 shades of blue) or more radical overhauls tend to yield more successful tests. Clearly, radical changes tend to be more successful. This may be because they are more likely to have a clear hypothesis behind them, while incremental changes are more speculative."

Even if a single A/B experiment is assessing an incremental idea, an experimentation strategy using A/B testing can lead to incremental and significant changes in the organization. Indeed, this argument is supported by the analysis of thousands of A/B tests by Qubit, an A/B testing consultancy. Its study finds that, although most tests yield small results, these aggregate into significant gains, and sometimes even a single well-crafted test produces a dramatic improvement (Browne and Jones 2017).

Other examples support the idea that firms use A/B testing for strategic decisions. One hardware manufacturer uses A/B tests to learn how responsive its customers are to purchasing its products from different retailers.[13] By testing which products sell best through which retailer, the manufacturer can structure its value chain and negotiate more effective contracts with its resellers. A digital media company has used A/B testing to better understand its evolving customer base, learning which products to introduce or retire. Finally, ride-sharing platforms have used A/B testing to learn how drivers and riders respond to new products and incentives that have the potential to cannibalize existing business revenue substantially.[14] These examples and the preceding qualitative evidence support the argument that A/B tests improve organizational learning by lowering the risk and speeding up the execution of strategic changes.

Finally, these insights also shed light on why relatively few start-ups adopt A/B testing tools. The quotes show that effective usage of A/B testing tools requires more than installing a few lines of code and randomizing the color of a button. Instead, when start-ups use A/B testing tools, they appear to shift how they generate ideas and decide between them. Similar to the arguments of Gans et al. (2019), start-ups appear to make costly commitments in order to benefit from low-cost experiments. Indeed, the adoption results in Table 1 suggest that such investments are more challenging for smaller, nonfunded, early stage teams to undertake compared with larger firms with more resources.

5.2. Measuring the Relationship Between A/B Testing and Significant Product Changes

To gain quantitative insight into whether firms in our data set are making significant shifts, we explore if A/B testing leads them to make more website and product changes. We used from the Internet Archive's Wayback Machine to extract monthly snapshots of the code that generates each start-up's home page. We focus only on start-ups that had raised funding at the beginning of our panel as they had more traction, generally received more public attention, and are more likely to be regularly indexed by the Wayback Machine. We then differenced these snapshots so that each observation in our data represents the difference between the current month and the next snapshot. In total, we have 8,268 start-ups with multiple observations. Using this data, we test if firms using A/B testing in a month t vary in how much they change their code before the next snapshot in t'. We fit a model of the form:

$$\Delta_{it,t'} = \beta(A/B\,Testing_{it}) + \theta(Technology\,Stack_{it}) + \eta(Log(Lines\,of\,Code_{it})) + \alpha_i + \gamma_t + \mu_{t-t'} + \epsilon_{it}, \quad (2)$$

where Δ_{it} is a measure of how much the website code changes (e.g., number of lines that are different) between t (the current month) and t'(the month of the next snapshot). The model includes firm and time fixed effects, and our technology stack control is calculated at the monthly level. We also include a control for website size Log(Lines of Code) and fixed effects for the number of months between snapshots to account for larger websites, and longer durations between snapshots feature more changes. However, our pattern of results does not shift if we exclude these two controls. We then calculate four different measures to evaluate changes to websites.

Log(Lines of code changed +1) is the total number of code lines that changed between t and t'. Although this is a coarse measure, changes to a code base have proved to be a useful proxy for how much a website or digital product has shifted (MacCormack et al. 2006).

Major code change (Top 5%) is a dichotomized version of our measure of lines changed to test whether the change is in the top 5% of lines of code changed. This measure allows us to test whether A/B testing firms make incremental changes to their code or more significant changes to their online presence.

Relative change in HTML structure captures differences in the underlying code tree. We use an algorithm developed by Gowda and Mattmann (2016) that compares HTML tree structures returning a dissimilarity score that that ranges from zero (the same HTML tree) to one (completely different). This measure allows us to determine whether A/B testing is associated with small tweaks in copy or pricing or more significant changes to the structure of a website.

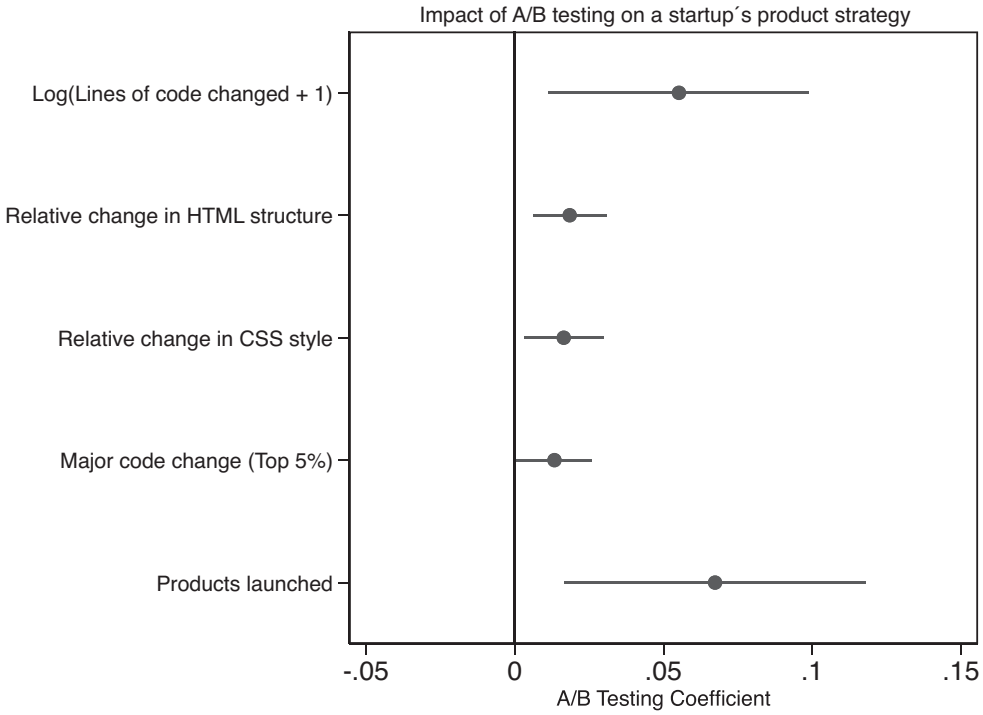
Relative change in Cascading Style Sheets (CSS) style uses a related algorithm that compares the similarity in the website's CSS to measure if A/B testing leads to incrementalism in design decisions. Again, this measure reflects a dissimilarity score ranging from zero (the same CSS style) to one (completely different). This measure allows us to test if A/B testing is more or less likely to change the visual presentation of a firm's product.

Although the Wayback Machine allows us to test if A/B testing changes how firms shift their code, it is an imperfect measure of a start-up's product development process. We undertake a second analysis to gain further insight into how A/B testing influences product development. We measure the number of products a start-up has launched during our panel by analyzing news coverage data collected by CrunchBase. As CrunchBase rarely links to news coverage for nonfunded start-ups, we focus our analysis here on the 13,186 start-ups that had raised funding at the start of our panel.

Products launched is calculated by parsing the titles in this set of Crunchbase-linked articles for each start-up for the strings "Introduce" or "Launch." Examples of article titles include "Madison Reed: Internet Retailer – Madison Reed launches an artificial intelligence chatbot," "Coinbase Launches OTC Platform, Clients Still Bullish on Crypto," and "Careem Introduces Credit Transfer." See Online Appendix A10 for additional examples of articles that did and did not get tagged as covering product launches. Because multiple reports might cover the same product launch, we count a product launch as a week with at least one of these articles. We use this variable to proxy whether the start-up is introducing new products. Because all these measures are at the start-up–week level, we use our basic model to test if A/B testing improves these metrics.

Figure 4 shows how A/B testing impacts the start-up's product development process.[15] The first result is the estimate on the number of lines changed between each Wayback Machine snapshot. This estimate suggests firms that adopt A/B testing change roughly 6% more lines of code than firms that do not use A/B testing. The second and third estimates show how different the HTML code structure and the website style are between snapshots. Again, we find positive estimates suggesting that A/B testing firms shift their code and CSS somewhat more aggressively than nonexperimenting firms. The fourth row shows that A/B testing increases the probability of a major code change. The final estimate suggests that A/B testing firms launch more new products as measured by CrunchBase news articles. On average, a start-up that uses A/B testing launches an additional 0.067 products per week. At the end of our product launch panel, the average firm had launched 0.36 products. Our estimate implies that an

Figure 4. We Find That A/B Testing Does Not Lead to Incrementalism in Product and Website Development for the Nearly 10,000 Start-ups for Which We Have Website and Product Launch Data



Notes. Instead, these firms make larger changes to their website code, the structure of their homepage's HTML, and website style and are more likely to deploy major code changes. A/B testing firms are also more likely to launch a new product in a given week than those that do not.

A/B testing firm has a risk of launching a product in a given week that is 18.6% greater than the average firm.

In sum, our qualitative evidence, along with our two empirical analyses, indicate that A/B tests were used to test significant product changes. If firms are testing more dramatic changes to their product offerings, it is more plausible that A/B testing drives organizational learning and significantly better performance. However, there is an additional implication from this logic. Not all tests find positive effects. Indeed, many of these A/B tests likely yield negative results that lead firms to abandon big ideas. In the next section, we explore whether the changes we observe firms making also drive greater performance variability.

5.3. A/B Testing, Learning, and Performance Variability

If A/B testing leads to significant product changes in an organization, variability in performance should also increase. This logic is well aligned with a growing academic and practitioner literature arguing that an efficient start-up has two natural endpoints: rapidly scaling or failing fast (Yu 2020). These outcomes are generally preferable to stagnation or slow growth. Rapid scaling is often necessary for high-technology start-ups because low entry barriers allow competitors to grab market share and eventually overtake first movers. However, if start-ups cannot scale, entrepreneurs are often advised to fail fast by venture capitalists and incubators (Yu 2020). For entrepreneurs with high opportunity costs, pivoting to a new idea can be more efficient than persistence in a lost cause (Arora and Nandkumar 2011, Camuffo et al. 2020, Yu 2020).

A/B testing helps start-ups recognize toward which of the natural endpoints they are headed. Tests may reveal incontrovertible evidence that none of the start-up's ideas are high quality. Moreover, the various changes start-ups make may not yield measurable performance gains on important metrics, such as visits, clicks, or sales. Alternatively, experimentation could help a start-up unearth a promising feature or customer segment that can scale quickly (Azevedo et al. 2020). Armed with the data from a major A/B test, entrepreneurs can take decisive action about whether to persist or pivot to a more promising idea or company.

These dynamics suggest that, although start-ups that use A/B testing should see increases in average performance, they should also be more likely to experience tail outcomes: scaling or failing.

5.4. Measuring if A/B Testing Increases Performance Variability

To explore this idea further, we conduct tests to examine if A/B testing leads to both increased scaling

and increased failing. To test this prediction, we split our website visits measure into five discrete and mutually exclusive buckets: zero, 1–499, 500–4,999, 5,000–49,999, and 50,000 or more weekly visits. Approximately 10% of firm-week observations are zero, and about 10% are more than 50,000. If A/B testing improves learning, adopting firms should be more willing to abandon bad ideas, leading them to have a higher chance of zero weekly visits. Further, because learning helps start-ups iterate in their search for product-market fit, firms should be more likely to end up with 5,000 or more views and less likely to remain mired in mediocrity in the sub-5,000 visits range. As discussed, because these measures are at the start-up–week level, we use our primary estimation technique.

Figure 5 shows the estimated effect of A/B testing on whether a start-up sample is more likely to scale or fail. Here, we see a bimodal response to the adoption of A/B testing just as predicted. This result suggests that A/B testing firms in our sample may be learning faster in terms of both whether their idea has little promise and how to scale their idea if it has potential. Our bimodal finding is related to work on how start-up accelerators improve learning, leading to increased failing and scaling (e.g., Yu 2020).

In assessing how A/B tests enhance firm performance, our additional analyses suggest, perhaps counterintuitively, that start-ups are using A/B tests to evaluate and implement significant product changes. These tests imply not only increased performance on average, but also increased variation in performance. We find further evidence on patterns of scaling and failing that is consistent with this conjecture.

6. Discussion and Conclusion

What is the right strategy for start-ups? Recent work in strategic management identifies experimentation as the preferred framework for decision making in young businesses (Gans et al. 2019, Camuffo et al. 2020). We exploit a technological shift in the cost of testing new ideas, enabled by the emergence of A/B testing software, to evaluate whether and how experimentation impacts start-up performance. We build a unique data set of more than 35,000 global start-ups, their adoption of A/B testing, and weekly performance measures. We find that A/B testing leads to a 10% increase in visits in the first few months after adoption for the start-ups in our sample. After a year of experimentation, the gains range from 30% to 100%. However, we find little evidence that the benefits of A/B testing vary between different kinds of start-ups. The most pronounced difference is between earlier and later stage start-ups with early stage start-ups appearing to benefit slightly more.

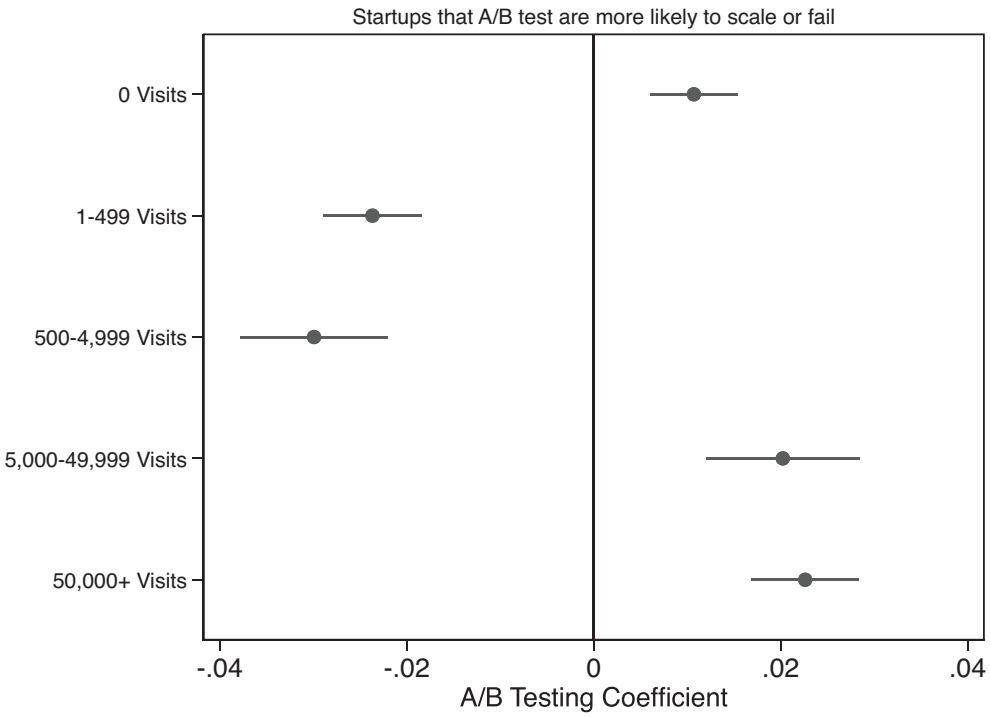
To explain these considerable performance improvements, we provide qualitative and quantitative evidence on how firms use A/B testing. Insights from practitioners suggest that A/B testing requires firms to shift their routines to realize the gains from experimentation. To complement this qualitative evidence, we demonstrate that firms using A/B testing launch more new products and make more significant changes to their website code. Further, these firms are more likely to scale or fail, a performance pattern consistent with the A/B testing of consequential ideas.

Our article informs two research agendas at the intersection of strategy and entrepreneurship. First, although our field has generated many insights about strategy in large organizations, we have only recently sought to clarify entrepreneurial strategy. Our findings provide empirical evidence that an experimental approach to strategy, as suggested by Levinthal (2017), Gans et al. (2019), and Camuffo et al. (2020), leads to better performance for young firms. We offer a novel insight highlighting an essential distinction between running *an* experiment versus a strategy based on experimentation. Running a single test most likely leads to null or negative results because most ideas fail (Kohavi and Longbotham 2017, Kohavi and Thomke 2017). A strategy based on experimentation helps firms quickly learn which of their ideas will significantly improve performance.

These findings also indicate a central role for organizational learning in entrepreneurial strategy (Levinthal 2017). The critical tension in entrepreneurial strategy is often framed analogously to the classic debate between Porter (1980) and Mintzberg (1990) over the benefits of intentional versus emergent strategy. However, in the context of entrepreneurial strategy, our results suggest that this comparison could be misleading. Consistent with Levinthal (2017), we find that a middle ground exists between a top-down strategy focused on credible commitments and a bottom-up strategy based solely on responding to the environment. An entrepreneurial strategy based on developing logical hypotheses, testing them rigorously, and incorporating the results into future strategic choices leads to competitive advantage. Commitment to experimentation via A/B testing appears to yield both valuable refinement of existing ideas and the development of significant product changes that contribute to better performance.

If experimentation is such a valuable framework for entrepreneurial strategy, however, why do fewer than one out of five firms in our sample ever adopt A/B testing? It could be that firms require complementary capabilities to leverage A/B testing fully. Indeed, we find that firms with more employees located in Silicon Valley and with VC funding are more likely to adopt A/B testing tools. Further, much of the qualitative evidence referenced in our paper suggests that experimentation must be embraced by senior leaders and permeate the entire organization to impact performance.

Figure 5. For the 35,262 Crunchbase Start-ups in Our Primary Sample, We Find A/B Testing Firms Are More Likely to Fail (End up with Zero-Visit Weeks) and Scale (Achieve More than 50,000 Visits in a Week)



Notes. The growth in tail outcomes comes at the expense of experiencing middling growth outcomes. Estimates are from a regression similar to Table 2, Model 4, but with the dependent variable dichotomized to reflect the visit ranges listed.

Thus, implementing this strategy in more developed organizations—with management teams, customers, and investors—is more complicated than in the case of lone entrepreneurs testing their minimum viable product (cf. Felin et al. 2019, Camuffo et al. 2020, Bennett and Chatterji 2022).

The second contribution of our work is to the emerging literature on data-driven decision making and the broader digitization of the global economy (Brynjolfsson et al. 2011, Brynjolfsson and McElheran 2016). This literature argues that firms' vast amount of transaction data allows them to do an unprecedented analysis of consumer data to inform their strategies. We demonstrate that A/B testing enables firms to do more than analyze the past. By generating, testing, and implementing *new* ideas, firms can use digital experimentation to design the future.

Our approach is not without limitations. Although we build the first large-panel data set on start-up experimentation, we recognize that A/B testing is not randomly assigned to firms. This selection challenge could bias our estimates upward although we take care to control important observed and unobserved factors that might drive A/B testing adoption and performance. Indeed, we show that our findings are robust when we use different identification strategies and methods ranging from instrumental variables to synthetic controls to difference-in-differences style models. In all our specifications, our results remain significant and consistent. Our effect sizes for technology start-ups, demonstrated on multiple metrics, are also in line with previous studies estimating the effect of data-driven decision making in large, publicly traded firms (Brynjolfsson et al. 2011).

Another challenge concerns generalizability. Our sample comprises start-ups competing in digital markets, which can experiment at a low cost. However, beyond digital markets, the cost of experimentation likely varies widely across industries. These industry differences are helpful in comparing our conclusions to those of other recent studies. For example, Gans et al. (2019) describe an entrepreneurial strategy based on experimentation requiring some level of commitment that forecloses alternative strategic choices. The firms in our sample can likely experiment via A/B testing with fewer commitments than in manufacturing or life sciences, in which experimentation may be costly (cf. Pillai et al. 2020). Future research can compare the impact of A/B testing on firms in these industries with other kinds of experimentation that require more significant trade-offs in the spirit of the Gans et al. (2019) "paradox of entrepreneurship."

Further, although we consider the long-term impact of A/B testing, we cannot evaluate how the adoption

of A/B testing influences intrafirm dynamics. We conjecture that these tools shape the design of organizations and roles as entrepreneurs seek to manage idea generation and implementation in new ways. Future research should investigate this phenomenon more deeply to better understand which organizational structures are most aligned with an experimental strategy. Finally, we do not observe the actual A/B tests that start-ups run, so our findings cannot discern whether A/B testing is part of a broader research and development program.

The continued decline in the cost of running digital experiments raises important questions for scholars and practitioners. How should managers design organizations that balance the flexibility enabled by experimentation with the reliable routines needed to execute? Moreover, although relatively few firms currently run digital experiments, will widespread adoption alter the benefits to individual organizations? Finally, how will experimentation across the economy change the types of innovations that firms develop and how they are distributed (e.g., Kerr et al. 2014, Cao et al. 2021)? Addressing these questions will guide future research and practice.

Acknowledgments

The authors thank seminar participants at the Harvard Business School, the Conference on Digital Experimentation at the Massachusetts Institute of Technology, Duke University, University of Maryland, Binghamton University, University of Minnesota, New York University, The Utah Winter Strategy Conference, the Strategy Science Conference, the Innovation Growth Laboratory Workshop, and the Wharton School for their feedback. The authors thank the Kauffman Foundation for their generous support of this work. Authors' names are in reverse alphabetical order. All authors contributed equally to this project.

Endnotes

[1] See https://www.optimizely.com/customers/blue-apron/.

[2] See https://engineering.upside.com/upside-engineering-blog-13-testing-culture-d5a1b659665e.

[3] One limitation of this approach is that acquired websites (e.g., Instagram or Bonobos) are not linked to their acquirers (e.g., Facebook, Walmart). That said, our results are robust to dropping firms marked as acquired in the Crunchbase data set. Further, our interest lies in how a start-up develops and grows its product(s), not corporate structures. For these reasons, we treat each start-up URL as an independent firm.

[4] In this scenario, a Facebook page, www.facebook.com/my-awesome-startup, would also be credited with Facebook's global page view numbers.

[5] There exists a much larger set of tools that have analytic capabilities or offer integration with A/B testing tools. We focus on tools that explicitly focus on A/B testing of a web application. Other tools, such as Mixpanel, are primarily analytics measurement tools. Although they integrate with A/B testing tools, using them does not necessarily indicate that a firm is running A/B tests. In this way, our estimates are conservative because firms in our counterfactual group may have adopted A/B testing and are labeled as not doing so.

[6] Online Appendix A2 presents these results as regression tables.

[7] In the TWFE models, the never- and always-users are used to estimate the week fixed effects and the technology stack control variable.

[8] In Online Appendix A9, we show the growth trajectory of six start-ups and the estimated counterfactual for that start-up. The model appears to capture varied growth trajectories at an appropriate level of detail.

[9] See https://auth0.com/blog/why-you-should-ab-test-everything/.

[10] See https://www.forbes.com/sites/quora/2013/07/01/how-do-designers-at-twitter-facebook-etc-balance-their-design-instinct-vs-the-engineers-impulses-to-use-ab-test-data-for-every-design-change/#3725037179a8.

[11] This was acquired by Optimizely.

[12] See https://priceonomics.com/optimizing-the-internet-what-kind-of-ab-testing/.

[13] See https://www.optimizely.com/customers/hp/.

[14] See https://eng.uber.com/xp/.

[15] In Online Appendix A11, we report the regression models that correspond to Figure 4. We also report models testing if A/B testing improves additional measures of product development. Again, we find positive effects.

References

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* 105(490):493–505.

Aral S, Walker D (2011) Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Sci.* 57(9):1623–1639.

Aral S, Walker D (2014) Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Sci.* 60(6):1352–1370.

Arora A, Nandkumar A (2011) Cash-out or flameout! Opportunity cost and entrepreneurial strategy: Theory, and evidence from the information security industry. *Management Sci.* 57(10): 1844–1860.

Arrow KJ (1962) The economic implications of learning by doing. *Rev. Econom. Stud.* 29(3):155–173.

Azevedo EM, Deng A, Olea JM, Rao JM, Weyl EG (2020) A/B testing with fat tails. *J. Political Econom.* 128(12).

Bapna R, Ramaprasad J, Shmueli G, Umyarov A (2016) One-way mirrors in online dating: A randomized field experiment. *Management Sci.* 62(11):3100–3122.

Bennett VM, Chatterji AK (2022) The entrepreneurial process: Evidence from a nationally representative survey. *Strategic Management J.* Forthcoming.

Bhidé A (1986) Hustle as strategy. *Harvard Bus. Rev.* 64(5):59–65.

Bhidé AV (2003) *The Origin and Evolution of New Businesses* (Oxford University Press, Oxford, UK).

Blank S (2013) *The Four Steps to the Epiphany: Successful Strategies for Products That Win* (John Wiley & Sons, Hoboken, NJ).

Boulding W, Lee E, Staelin R (1994) Mastering the mix: Do advertising, promotion, and sales force activities lead to differentiation? *J. Marketing Res.* 31(2):159–172.

Browne W, Jones MS (2017) What works in e-commerce: A meta-analysis of 6700 online experiments. [White Paper]. Qubit Digital Ltd. Accessed January 4, 2022, https://fs.hubspotusercontent00.net/hubfs/215600/Qubit%20meta%20analysis%20%5Bacademic-paper%5D.pdf.

Brynjolfsson E, Hitt LM (2003) Computing productivity: Firm-level evidence. *Rev. Econom. Statist.* 85(4):793–808.

Brynjolfsson E, McAfee A (2012) *Race against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy* (Digital Frontier Press, Lexington, MA).

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

Brynjolfsson E, McElherfan KS (2019) Data in action: Data-driven decision making and predictive analytics in US manufacturing. Preprint, submitted July 19, https://ssrn.com/abstract=1819486" https://ssrn.com/abstract_id=3422397.

Brynjolfsson E, Hitt LM, Kim HH (2011) Strength in numbers: How does data-driven decision making affect firm performance? Preprint, submitted April 22, https://ssrn.com/abstract=1819486.

Camerer C, Lovallo D (1999) Overconfidence and excess entry: An experimental approach. *Amer. Econom. Rev.* 89(1):306–318.

Camuffo A, Cordova A, Gambardella A, Spina C (2020) A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Sci.* 66(2):564–586.

Cao R, Koning R, Nanda R (2021) Biased sampling of early users and the direction of startup innovation. Preprint, submitted June, https://www.nber.org/papers/w28882.

Chatterji AK, Fabrizio KR (2014) Using users: When does external knowledge enhance corporate product innovation? *Strategic Management J.* 35(10):1427–1445.

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

Cohen WM, Levinthal DA (1994) Fortune favors the prepared firm. *Management Sci.* 40(2):227–251.

Cohen WM, Nelson RR, Walsh JP (2002) Links and impacts: The influence of public research on industrial R&D. *Management Sci.* 48(1):1–23.

Dahlander L, Piezunka H (2014) Open to suggestions: How organizations elicit suggestions through proactive and reactive attention. *Res. Policy* 43(5):812–827.

David PA (1975) *Technical Choice Innovation and Economic Growth: Essays on American and British Experience in the Nineteenth Century* (Cambridge University Press, Cambridge, UK).

Deniz BC (2021) Experimentation and incrementalism: The impact of the adoption of A/B Testing. Preprint, submitted February, https://api.semanticscholar.org/CorpusID:235375442.

Denrell J (2003) Vicarious learning, undersampling of failure, and the myths of management. *Organ. Sci.* 14(3):227–243.

Denrell J, March JG (2001) Adaptation as information restriction: The hot stove effect. *Organ. Sci.* 12(5):523–538.

Dubé J-P, Fang Z, Fong N, Luo X (2017) Competitive price targeting with smartphone coupons. *Marketing Sci.* 36(6):944–975.

Dyer JH, Hatch NW (2004) Using supplier networks to learn faster. *MIT Sloan Management Rev.* 45(3):57.

Fabijan A, Dmitriev P, Olsson HH, Bosch J (2017) The evolution of continuous experimentation in software product development: From data to a data-driven organization at scale. *Proc. 39th Internat. Conf. Software Engrg.* (IEEE Press, Piscataway, NJ), 770–780.

Felin T, Gambardella A, Stern S, Zenger T (2019) Lean startup and the business model: Experimentation revisited. *Long Range Planning* 53(4):101953.

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

Ghemawat P (1991) *Commitment* (Simon and Schuster, New York).

Ghemawat P, Del Sol P (1998) Commitment versus flexibility? *California Management Rev.* 40(4):26–42.

Gomez-Uribe CA, Hunt N (2016) The Netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Management Inform. Systems* 6(4):1–13.

Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. *J. Econometrics* 225(2):254–277.

Gowda T, Mattmann CA (2016) Clustering web pages based on structure and style similarity (application paper). *2016 IEEE 17th Internat. Conf. Inform. Reuse Integration* (IEEE), 175–180.

Hannan M (1984) Structural inertia and organizational change. *Amer. Sociol. Rev.* 49(2):149–164.

Hasan S, Koning R (2019) Prior ties and the limits of peer effects on startup team performance. *Strategic Management J.* 40(9):1394–1416.

Hendel I, Spiegel Y (2014) Small steps for workers, a giant leap for productivity. *Amer. Econom. J. Appl. Econom.* 6(1):73–90.

Jeppesen LB, Lakhani KR (2010) Marginality and problem-solving effectiveness in broadcast search. *Organ. Sci.* 21(5):1016–1033.

Kerr WR, Nanda R, Rhodes-Kropf M (2014) Entrepreneurship as experimentation. *J. Econom. Perspect.* 28(3):25–48.

King AA, Goldfarb B, Simcoe T (2019) Learning from testimony on quantitative research in management. *Acad. Management Rev.* 46(3):465–488.

Knudsen T, Levinthal DA (2007) Two faces of search: Alternative generation and alternative evaluation. *Organ. Sci.* 18(1):39–54.

Kohavi R, Longbotham R (2017) Online controlled experiments and A/B testing. *Encyclopedia of Machine Learning and Data Mining* 7(8):922–929.

Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard Bus. Rev.* 95(5).

Kohavi R, Henne RM, Sommerfield D (2007) Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO. *Proc. 13th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM), 959–967.

Kohavi R, Crook T, Longbotham R, Frasca B, Henne R, Ferres JL, Melamed T (2009) Online experimentation at Microsoft. Data Mining Case Studies 11. Accessed January 4, 2022, https://ai.stanford.edu/~ronnyk/ExPThinkWeek2009Public.pdf.

Kumar A, Tan Y (2015) The demand effects of joint product advertising in online videos. *Management Sci.* 61(8):1921–1937.

Lawrence A, Ryans J, Sun E, Laptev N (2018) Earnings announcement promotions: A Yahoo Finance field experiment. *J. Accounting Econom.* 66(2–3):399–414.

Levinthal DA (2017) Mendel in the C-suite: Design and the evolution of strategies. *Strategy Sci.* 2(4):282–287.

Levitt SD, List JA, Syverson C (2013) Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *J. Political Econom.* 121(4):643–681.

Luca M, Bazerman MH (2021) *The Power of Experiments: Decision Making in a Data-Driven World* (MIT Press, Cambridge, MA).

MacCormack A, Rusnak J, Baldwin CY (2006) Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Sci.* 52(7):1015–1030.

Madsen PM, Desai V (2010) Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. *Acad. Management J.* 53(3):451–476.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

McDonald RM, Eisenhardt KM (2020) Parallel play: Startups, nascent markets, and effective business-model design. *Admin. Sci. Quart.* 65(2):483–523.

McGrath RG (1999) Falling forward: Real options reasoning and entrepreneurial failure. *Acad. Management Rev.* 24(1):13–30.

McGrath RG, MacMillan IC (2000) The entrepreneurial mindset: Strategies for continuously creating opportunity in an age of uncertainty, vol. 284 (Harvard Business Press, Cambridge, MA).

McMullen JS, Shepherd DA (2006) Entrepreneurial action and the role of uncertainty in the theory of the entrepreneur. *Acad. Management Rev.* 31(1):132–152.

Mintzberg H (1990) The design school: Reconsidering the basic premises of strategic management. *Strategic Management J.* 11(3):171–195.

Mowery DC, Oxley JE, Silverman BS (1996) Strategic alliances and interfirm knowledge transfer. *Strategic Management J.* 17(S2):77–91.

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. General Psych.* 2(2):175–220.

Nordhaus WD (2007) Two centuries of productivity growth in computing. *J. Econom. History* 67(1):128–159.

Ott TE, Eisenhardt KM (2020) Decision weaving: Forming novel, complex strategy in entrepreneurial settings. *Strategic Management J.* 41(2):2275–2314.

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 J.* 41(3): 369–399.

Porter ME (1980) *Competitive Strategy: Techniques for Analyzing Industries and Competitors* (Free Press, New York).

Ries E (2011) *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (Crown Books, New York).

Runge J, Nair H (2021) Exploration in action: The role of randomized control trials in online demand generation. Working paper, Stanford GSB, CA.

Sahni NS, Zou D, Chintagunta PK (2016) Do targeted discount offers serve as advertising? Evidence from 70 field experiments. *Management Sci.* 63(8):2688–2705.

Sarasvathy SD (2001) Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. *Acad. Management Rev.* 26(2):243–263.

Simon HA (1959) Theories of decision-making in economics and behavioral science. *Amer. Econom. Rev.* 49(3):253–283.

Siroker D, Koomen P, Kim E, Siroker E (2014) U.S. Patent 8,839,093. (U.S. Patent and Trademark Office, Washington, DC).

Sitkin SB (1992) Learning through failure: The strategy of small losses. *Res. Organ. Behav.* 14:231–266.

Thomke S (2001) Enlightened experimentation: The new imperative for innovation. *Harvard Bus. Rev.* 79(2):66–75.

Thomke SH (2020) *Experimentation Works: The Surprising Power of Business Experiments* (Harvard Business Review Press, Boston).

Timmermans S, Tavory I (2012) Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociol. Theory* 30(3):167–186.

Urban GL, Katz GM (1983) Pre-test-market models: Validation and managerial implications. *J. Marketing Res.* 20(3):221–234.

Urban GL, Von Hippel E (1988) Lead user analyses for the development of new industrial products. *Management Sci.* 34(5):569–582.

Van den Steen E (2016) A formal theory of strategy. *Management Sci.* 63(8):2616–2636.

Xu Y (2015) Why experimentation is so important for LinkedIn. Accessed January 4, 2022, https://engineering.linkedin.com/ab-testing/why-experimentation-so-important-linkedin.

Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Anal.* 25(1):57–76.

Xu Y, Chen N, Fernandez A, Sinno O, Bhasin A (2015) From infrastructure to culture: A/B testing challenges in large scale social networks. *Proc. 21st ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM), 2227–2236.

Yu S (2020) How do accelerators impact the performance of high-technology ventures? *Management Sci.* 66(2):530–552.