HARVARD | BUSINESS | SCHOOL

Experimentation and Startup Performance: Evidence from A/B Testing

Rembrand Koning
Sharique Hasan
Aaron Chatterji

Working Paper 20-018

Experimentation and Startup Performance: Evidence from A/B Testing

Rembrand Koning
Harvard Business School

Sharique Hasan
Fuqua School of Business, Duke University

Aaron Chatterji
Fuqua School of Business, Duke University and
NBER

Working Paper 20-018

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

Experimentation and startup performance: Evidence from A/B testing*

Rembrand Koning
Harvard Business School
rem@hbs.edu

Sharique Hasan
Duke Fuqua
sh424@duke.edu

Aaron Chatterji
Duke Fuqua and NBER
ronnie@duke.edu

August 20, 2019

Abstract

Recent work argues that experimentation is the appropriate framework for entrepreneurial strategy. We investigate this proposition by exploiting the time-varying adoption of A/B testing technology, which has drastically reduced the cost of experimentally testing business ideas. This paper provides the first evidence of how digital experimentation affects the performance of a large sample of high-technology startups using data that tracks their growth, technology use, and product launches. We find that, despite its prominence in the business press, relatively few firms have adopted A/B testing. However, among those that do, we find increased performance on several critical dimensions, including page views and new product features. Furthermore, A/B testing is positively related to tail outcomes, with younger ventures failing faster and older firms being more likely to scale. Firms with experienced managers also derive more benefits from A/B testing. Our results inform the emerging literature on entrepreneurial strategy and how digitization and data-driven decision-making are shaping strategy.

*Authors names are in reverse alphabetical order. All authors contributed equally to this project. We thank seminar participants at Harvard Business School, the Conference on Digital Experimentation, Duke, University of Maryland, Binghamton University, University of Minnesota, NYU, and Wharton for their feedback. We thank the Kauffman Foundation for their generous support of this work.

Introduction

Why do so few startups succeed? For mature companies, scholars often attribute success and failure 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 startups need a more flexible strategic framework. Levinthal (2017) articulates an experimental approach to entrepreneurial strategy and emphasizes the role of a "Mendelian" executive who generates alternatives, tests their efficacy, and selects the best course. Similarly, Camuffo et al. (2019) advise entrepreneurs to propose and test many hypotheses about their startup's strategy. Gans, Stern and Wu (2019) also advocate experimentation but underline the importance of commitment when choosing among 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), implementing one has historically been costly. In recent years, however, the cost of running many experiments has declined due to the digitization of the economy and the proliferation of A/B testing tools (Kohavi, Henne and Sommerfield, 2007; Kohavi et al., 2009; Kohavi and Longbotham, 2017; Azevedo et al., 2018). With this technology, firms of all sizes and vintages can now rapidly implement many experiments to test business decisions. Accelerators, venture capital firms, and leading entrepreneurs advocate that startups should A/B test everything. But does A/B testing alone constitute an experimental strategy that helps startups succeed?

While A/B testing has reduced the cost of evaluating competing ideas, experimentation is a much broader concept. Experimentation comprises generating alternatives, testing them, and selecting the most appropriate solution (Levinthal, 2017; Camuffo et al., 2019; Gans, Stern and Wu, 2019). For A/B testing adoption to matter for performance, firms must also continue to generate new ideas and be willing to let data drive

decisions (Levinthal, 2017). The availability of low-cost testing tools could incentivize idea generation and spark a culture of data-driven decision-making (Brynjolfsson and McElheran, 2016).

But will all firms realize the benefits of such experimentation? Given the crucial role of idea generation and selection, some firms may execute an experimental strategy better than others (Levinthal, 2017). Firms with more experienced managers may generate better ideas, resulting in better options to choose from. Moreover, firms with more focused product offerings will be better able to decide among equally viable alternatives (Gans, Stern and Wu, 2019).

We evaluate whether and how experimentation enhances startup performance. Historically, this research question has been challenging to address empirically because accurate measurement of economy-wide experimentation has been prohibitive. We overcome this hurdle by combining three heretofore distinct sources of data to examine the impact of the adoption of A/B testing in 35,000 global startups. Our data combines a panel of consistent measures of when firms adopt A/B testing with weekly measures of their performance between 2015 and 2018. We complement this data with a rich set of information about each startup's technology stack, funding, product characteristics, and management team.

First, we find considerable heterogeneity in when firms adopt A/B testing technology. At the start of our panel, only 7.6% of startups use A/B testing technology, whereas over time, 16.75% of startups in our sample adopt the technology. In a variety of demanding specifications, we find that A/B testing is associated with a persistent 11% increase in page views. Additional tests support the theory that while A/B testing in practical terms only reduces the costs of conducting tests, the technology appears to affect both the rate at which startups introduce new products and the heterogeneity of their outcomes. Firms that adopt A/B testing introduce new products at a 9% to 18% higher rate than those who do not experiment. Furthermore, we find evidence that this technology is associated with an increased likelihood of tail outcomes—i.e., more zero page-view weeks and more 100k+ page view weeks. Additional analysis suggests

that younger startups ‘ ‘fail faster” when they A/B test—with a persistently increased likelihood of a website that receives no page views.

However, not all startups can capitalize on this technology equally. Startups with more experienced managers—those who have deep expertise in a product category—appear to benefit most from A/B testing. Our results are consistent with prior theory (Levinthal, 2017; Camuffo et al., 2019; Gans, Stern and Wu, 2019). Implementing an experimental strategy will require complementary capabilities, not just cheaper testing.

Our findings contribute to the literature in at least three important respects. First, we contribute to a growing literature on entrepreneurial strategy by providing large-scale empirical support that a flexible and experimental approach to strategy can lead to persistent performance improvements, but that firms must also have the necessary capability to profit from experimentation (Camuffo et al., 2019; Gans, Stern and Wu, 2019; Levinthal, 2017). Second, despite the benefits of A/B testing and the growing consensus that firms should experiment (Xu et al., 2015; Kohavi et al., 2009; Thomke, 2001; March, 1991), few firms do. This finding is striking. Even young high-technology companies experiment infrequently, despite the fact that the cost of doing so has declined precipitously. Finally, we contribute to the literature on data-driven decision-making in the economy (Brynjolfsson and McElheran, 2016), in that A/B testing allows firms to base strategic decisions on a rigorous scientific process, not just intuition (Camuffo et al., 2019).

Theoretical framework

Entrepreneurial strategy and experimentation

Uncertainty is endemic to the entrepreneurial process (McMullen and Shepherd, 2006). Entrepreneurs must make many decisions, often with uncertain or unknown payoffs. They must choose which customers to serve, what product features to include, or 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 experimentation might be the right approach for entrepreneurial strategy (Levinthal, 2017; Camuffo et al., 2019; Gans, Stern and Wu, 2019). In this work, experimentation is cast as a three-part framework. 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 flexibility, avoiding premature or costly commitments (Bhide, 1986; Bhidé, 2003).

While experimentation has long been promoted as a framework for strategic decisions by academics (Thomke, 2001; Bhidé, 2003) and practitioners (Kohavi and Longbotham, 2017; Blank, 2013; Ries, 2011), 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. Finally, deciding among viable options is also challenging (Simon, 1959). Bureaucracy and other sources of inertia inside organizations may hinder the ability to take decisive action (e.g., Hannan, 1984). For these reasons and others, firms have rarely used formal experiments to inform business decisions.

In the last decade, rapid digitization of the global economy has altered this calculus (Brynjolfsson and McAfee, 2012). In particular, the cost of actually running controlled tests that compare alternatives has declined dramatically (Kohavi, Henne and Sommerfield, 2007; Kohavi and Longbotham, 2017). This is because 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., 2018). Today, software products like Optimizely and Google Optimize allow any firm with a digital presence to set up controlled experiments and analyze the data using prepackaged software.

What does A/B testing technology do?

Although no prior studies have examined the benefits of A/B testing across many firms, practitioners have written publicly about the utility of experimentation in their own organizations (Kohavi, Henne and Sommerfield, 2007; Kohavi et al., 2009; Kohavi and Longbotham, 2017; Xu et al., 2015). LinkedIn, the professional social network owned by Microsoft, sees A/B testing as vital to its business strategy. Xu (2015), who leads the company's core engineering team, describes the link between its use of digital experimentation technology and the company's strategy:

> *As LinkedIn grows, we increasingly rely on good strategies to take our products to the next level, so we focus our time and resources on ideas that give us the best return on investment. A/B testing offers the most scientific approach to assess the impact of any change and to draw clear causal relationships.*

Netflix, the streaming video service, also views A/B testing as central to its product innovation success. Urban, Sreenivasan and Kannan (2016) describe Netflix's use of A/B testing for all product decisions, large or small:

> *Ever wonder how Netflix serves a great streaming experience with high-quality video and minimal playback interruptions? Thank the team of engineers and data scientists who constantly A/B test their innovations to our adaptive streaming and content delivery network algorithms. What about more obvious changes, such as the complete redesign of our UI layout or our new personalized homepage? Yes, all thoroughly A/B tested.*
>
> *In fact, every product change Netflix considers goes through a rigorous A/B testing process before becoming the default user experience.*

Firms use A/B testing to evaluate new ideas that affect many aspects of their business (Xu et al., 2015; Kohavi, Henne and Sommerfield, 2007). Online retailers, for

example, test different bundling, pricing, and product display strategies (Dubé et al., 2017; Sahni, Zou and Chintagunta, 2016). Networking platforms experiment with social features, recommendation algorithms, or content to increase user engagement (Bapna et al., 2016; Kumar and Tan, 2015; Aral and Walker, 2014, 2011). Media companies A/B test the placements of articles or videos on their website, title variants, or subscription prices (Lawrence et al., 2018; Gomez-Uribe and Hunt, 2016).

Firms also experiment before they make critical strategic decisions. For example, a hardware manufacturer runs A/B tests to measure how responsive its customers are to purchasing its products from different retailers.[1] By testing which products sell best through which retailer, the hardware manufacturer can structure its value chain and negotiate more effective contracts with its resellers. A digital media company uses A/B testing to understand its evolving customer base better and use the results of its experiments to decide which new products to introduce and which ones to exit. Finally, a ride-sharing platform uses A/B testing to quickly learn about how drivers respond to different incentives when entering a new geographic market.[2]

These anecdotes may suggest that A/B testing and experimentation are equivalent in practice. However, a strategy based on experimentation requires more than just a reduction in the cost of testing ideas (Fabijan et al., 2017)—which is the key innovation facilitated by Optimizely and other A/B testing platforms (see, for example Siroker et al., 2014). Recall that experimentation has three parts, including the introduction of variation, the testing of alternatives, and the selection of candidate solutions (Camuffo et al., 2019). In the absence of many good ideas (Levinthal, 2017) and a rigorous decision-making process (Gans, Stern and Wu, 2019), A/B testing might not make a difference.

Alternatively, if the cost of running formal tests declines, firms may have stronger incentives to develop better ideas and implement them. Finally, some firms may have the complementary capabilities to benefit from the reduced cost of testing, such as

[1] https://www.optimizely.com/customers/hp/
[2] https://eng.uber.com/xp/

managerial experience and a well-defined product strategy. Below, we provide further detail on how A/B testing could affect firm performance.

Does A/B testing lead to more ideas?

Prior research suggests that when the cost of an important input declines, organizations often respond by investing in complementary practices (Nordhaus, 2007). For example, researchers have documented that information technology investments yielded returns for firms when they invested in hiring workers with relevant expertise and changing organizational routines (Brynjolfsson and Hitt, 2003). Likewise, the reduced cost of testing ideas may incentivize a firm to invest in idea generation. A/B testing encourages firms to introduce new and even incomplete products to a subset of users in a highly controlled environment, without affecting the product experience for the vast majority of users. Even after a new-product release, firms can A/B test to iteratively improve their products, per Xu (2015):

> *We rely on experimentation to guide product development not only because it validates or invalidates our hypotheses, but more importantly, because it helps create a mentality around building MVP and exploring the terrain around it.*
>
> *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.*

A/B testing may further encourage idea generation by facilitating objective arbitration between competing ideas. Prior research reveals that employees are a valuable source of knowledge. However, they are often cynical about bureaucracy and fear negative appraisals, which discourages idea generation (Girotra, Terwiesch and Ulrich, 2010). A/B testing allows individuals from any part of the organization to participate in a transparent idea-generation and -testing process. John Cline, Director of Engineering at Blue Apron, highlights how an A/B testing platform led to more product ideas:

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.

Scaling and failing

A growing academic and practitioner literature argues that an efficient startup has two natural endpoints: rapidly scaling or failing fast. These outcomes are preferable to stagnation or slow growth. Rapid scaling is often a necessity for high-technology startups because low barriers to entry allow competitors to grab market share and eventually overtake first movers. However, if startups cannot scale, entrepreneurs should fail fast. 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).

A/B testing helps startups recognize which of the natural endpoints they are headed toward. Since A/B testing increases the number of ideas generated, the variation in their quality also increases (Girotra, Terwiesch and Ulrich, 2010). With an influx of new ideas, good and bad, startups the challenge of selecting the very best grows as well. A/B testing also provides entrepreneurs with rigorous evidence on which of these ideas are drawn from the right tail of the quality distribution. Implementing these ideas over others will lead to rapid scaling. Moreover, as highlighted by Azevedo et al. (2018) and Kohavi and Thomke (2017), incremental iteration via A/B testing can hone these high-potential ideas even further.

Alternatively, A/B testing may reveal incontrovertible evidence that none of the startup's ideas are high-quality. Moreover, incremental changes may not yield measurable performance gains. Armed with this data, entrepreneurs can take decisive action about whether to persist or pivot to a more promising idea or company.

Given these two mechanisms, startups that use A/B testing should not only see increases in their average performance, but they should also be more likely to experience tail outcomes, scaling or failing.

Who benefits from A/B testing?

Startups vary considerably in their experience and capabilities. As a consequence, the mechanisms described above will operate differently depending on these characteristics.

To successfully implement an experimentation strategy, startups must generate many promising ideas to test. Prior research indicates that not all entrepreneurs are equally likely to do so. Entrepreneurs with prior industry experience have technical knowledge, market knowledge, and connections with customers that help them generate better ideas (Chatterji and Fabrizio, 2012; Gompers, Lerner and Scharfstein, 2005; Klepper and Sleeper, 2005; Agarwal et al., 2004; Shane, 2000). These entrepreneurs will benefit disproportionately from A/B testing.

Another key difference is whether a company has "product-market fit"—demonstrated customer demand and a workable business model—or is still searching for it (Ries, 2011). Older firms are more likely to have product-market fit and use A/B testing to make improvements on already successful products (Sørensen and Stuart, 2000). Young startups typically lack product-market fit and will use A/B testing to evaluate whether they should persist with or pivot from an initial business idea. The implication is that more mature startups that use A/B testing will scale quickly while newer ventures will fail faster.

In sum, we make the following predictions about how A/B testing aligns with the experimental strategy described in recent work (Levinthal, 2017; Gans, Stern and Wu, 2019; Camuffo et al., 2019). First, we predict that startups that implement A/B testing technology will see an increase in their performance. Moreover, if startups respond to the lower cost of A/B testing by increasing the number of ideas, they should also have more new product introductions. Further, if A/B testing also facilitates data-driven decisions, we should expect startups that implement this technology to scale and fail faster. We also expect the benefits of A/B testing to be more pronounced for firms with experienced managers. Finally, we predict that A/B testing will accelerate development toward rapid growth or pivoting, for older and younger startups, respectively.

What is the alternative to experimentation?

Before proceeding to our empirical approach, we consider the appropriate counterfactuals for A/B testing. If startups are not experimenting, how are they making decisions? Two approaches have been highlighted in prior work. First, a large 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 (McGrath, 1999; Nickerson, 1998). The implication is that an overconfident entrepreneur will invest time and effort into implementing strategies that will likely fail. This approach will drive performance differences between firms that experiment and those that do not.

However, a second literature documents that firms have long performed "uncontrolled experiments" or tweaks to their product development process (David et al., 1975). Hendel and Spiegel (2014) attribute much of the large productivity gains in a steel mill they studied over a 12-year period to learning from uncontrolled experiments. These tweaks include experimenting with how scrap enters the furnace and the timing of various production tasks. Levitt, List and Syverson (2013) document a similar phenomenon in a large automaker's assembly plant where learning by doing (Arrow, 1962) leads to productivity gains. In our sample of high-technology startups, it is possible that firms that are not conducting formal experiments are tweaking their products informally, which could lead to improved performance, albeit at a slower pace. Taken together, A/B testing should reduce the false positives of confirmatory search and accelerate the rate of discovering product improvements as compared to tweaking.

Data and Methods

Data sources

To test our predictions, we construct a longitudinal data set comprising 35,918 high-technology startups founded between 2008 and 2014. Our data include information

about these startups compiled from three distinct sources. Crunchbase provides us detailed information about each startup's product, funding status, and age as well as details about the founding team. We complement the Crunchbase information with weekly measures of page views for each startup from SimilarWeb and refer to the BuiltWith database for information on the technologies the startups use to build their product. Below, we describe the construction of our panel in detail.

Crunchbase Pro is a subscription database that tracks technology startups across the globe. The database is used primarily for lead generation, competitor analysis, and investment/acquisition research by industry users. Crunchbase's coverage of internet-focused startups is comparable to other startup data products (Kaplan and Lerner, 2016). While the database does include large technology companies such as Google and Microsoft, the majority of firms in its sample are startups. The quality of information about these startups improves significantly after 2008 and includes information on the startups including founding year, firm name, company website, funding raised, and a brief description of the startup's product. Crunchbase also provides links to news articles covering investments, product launches, and other key company events. Finally, CrunchBase also provides information on the founding team, executives, and board members. This data is particularly reliable for companies that have raised funding. [3]

Builtwith is a lead-generation, sales intelligence, and market share analysis platform for web technologies. Companies like Google, Facebook, and Amazon use this database to learn about the adoption of software components used to build web applications. The set of elements used to develop an application (e.g., database, back-end frameworks, front-end frameworks) are colloquially known as a "technology stack" BuiltWith indexes 30,000 web technologies across over 250 million websites. It tracks these websites' current and prior technology stacks. Figure 1 shows the data BuiltWith has on bombas.com, a direct-to-consumer apparel startup founded in 2013 that appears in Crunchbase startup data. BuiltWith tracks when each technology was first detected

[3]While the number of employees for each startup is reported on Crunchbase, only the most recent estimate is available.

and when it was uninstalled. Using the Crunchbase data as our base sample, we download each company's profile on BuiltWith to construct detailed technology adoption histories for the companies in our sample. Using this data, we can identify when a company adopts A/B testing technology into its stack.

SimilarWeb is a market intelligence platform that estimates website and app growth metrics. Using data from a global panel of web browsers, SimilarWeb provides website performance metrics including page views, bounce rates, and time on site over the last three years at the weekly level. SimilarWeb is used by firms like Airbnb, Procter & Gamble, and Deloitte for lead generation, to track acquisition targets, and to benchmark performance. We use the SimilarWeb API to pull down weekly website performance metrics for the Crunchbase startups in our sample.

Sample construction

We link startups across these three data sources through website URLs. Unlike firm names, URLs are unique identifiers, eliminating the need for fuzzy matches.[4] To ensure that our sample consists of "active" startups, we include only startups in the Crunchbase data that have non-zero page views in March 2015, the first month for which we have SimilarWeb data. We also exclude startups that have sub-domains—versus primary domains—as URLs, since SimilarWeb does not provide independent estimates for sub-domains.[5] Finally, some startups consist of thousands of sub-domains. In BuiltWith, for example, technologies used by sub-domains are attributed to the parent domain (e.g., wordpress.com would be assigned any technology attributed to my-awesome-blog.wordpress.com). To address this problem, we exclude pages with more than 80 active unique technologies as of March 2015.

[4]One limitation of this approach, however, 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 startup develops and grows its product(s), and not in the impact of corporate structures. For these reasons, we treat each startup URL as an independent firm.

[5]This means a Facebook page www.facebook.com/my-awesome-startup would get Facebook's global page view numbers.

After these exclusions, our dataset consists of 35,918 independent product-oriented startups founded between 2008 through 2013. Our panel captures the characteristics, web metrics, and technology adoption trajectories of these startups starting in the week of April 5th, 2015 until July 22nd, 2018 resulting in 173 weeks and a total of 6.2 million firm-week observations.

Estimation Strategy

Because our study is observational, we use several demanding econometric specifications to isolate the impact of A/B testing adoption—capturing the reduced cost of experimentation—on startup performance. To do so, we estimate equation 1 below:

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

To estimate the impact of A/B testing adoption (β) on startup performance (Y_{it}), we estimate a fixed effects model that isolates the change in firm outcomes before and after the adoption of A/B testing tools by a startup. We include fixed effects for each week (γ_t) to control for observed and unobserved non-parametric 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 of (β). 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 startup idea, the presence of a strategy, location, and the educational background of founders, among other fixed resources or capabilities of the startups.

In addition to our fixed effects, we include a weekly time-varying control for the number of other technologies adopted by the startup. Including this control allows us to account for time-varying changes to other elements of the startup's technological profile. Including this control increases our confidence that observed changes in performance attributed to A/B testing are not derived from other changes to the technology stack of a company (e.g., Google Analytics). Finally, in our most stringent specifications, we

also include fixed effects for the firm's growth rates—heterogenous firm-week slopes. Including these variables allows us to address the possibility that our finding is a consequence of fast-growing startups adopting A/B testing at higher rates.

Variable construction

Below we describe the main independent and dependent variables used in our study.

Independent variables

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 360, Google Website Optimizer, Omniture Adobe Test and Target, Optimization Robot, Optimizely, Optimost, Split Optimizer, and Visual Website Optimizer.[6]

Roughly 16.75% of firms use an A/B testing tool in the 173 weeks of our data. On average, 8.25% of firms actively use A/B testing technology in any given week. In our data, Optimizely is the market leader, accounting for 55% of the weeks in which firms are A/B testing. The next most prominent A/B testing software is Google, with just under 22% of the market. Finally, Visual Website Optimizer accounts for 18% of the market, and the remaining 5% is split between Adobe, AB Tasty, and Experiment.ly.

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 that were active on the website according to BuiltWith at the start of the week. Over the 173 weeks, we see some firms drop to four technologies (1st percentile) while others use over

[6]There exists a much large 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, like Mixpanel, are primarily analytics measurements tools, and while they integrate with A/B testing tools, using them does not necessarily indicate a firm is running A/B tests. In this way, our estimates are conservative, since firms in our counterfactual group may have adopted A/B testing as well, but are labeled as not doing so.

110 technology elements to build their web application (99th percentile). To account for the skewness in the technology adoption data, we log this variable. However, results are unchanged when we include the raw counts.

Age measures the firm's age in years. We interact age with whether the firm is using A/B testing to test if older and younger startups experience differential outcomes as a result of using A/B testing technology.

Experience is a count of the number of prior startups a firm's focal founder worked at previously. We interact experience with whether the firm is using A/B testing to test if startups with more experienced founders have differential outcomes as a result of using A/B testing technology.

Dependent variables

Our analysis examines the impact of A/B testing on three related startup performance metrics. These are described below.

Log(Pageviews+1) is the log of the weekly page views as estimated by SimilarWeb. Since page views can drop to 0 so we add 1 before transforming the variable.

Number of products launches counts the number of weeks where there is news coverage of the startup launching or introducing a new product or significant feature. We pulled news coverage data for 13,186 startups that had raised funding at the start of our panel, since CrunchBase has inadequate news coverage for non-funded startups. We calculate this variable by parsing the titles in the set of Crunchbase linked articles for each startup for "Introduces" or "Launches." 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." Since multiple articles might cover the same prod-

uct launch, we count a product launch as a week with at least one of these articles. We use this variable to proxy whether the startup is engaging in new-product idea generation.

0 pageview week is a dummy for whether the startup had a zero page view week. Consistent zero page view weeks serves as our proxy for an increase in the likelihood of startup failure. This allows us to test if A/B testing impacts the left tail, not just the mean, of startup outcomes.

100k+ pageview week is a dummy for whether the startup had more than 100,000 page views in a week, the 95th percentile of page views in our data. We use this longitudinal variable to test if A/B testing affects the right tail of the performance distribution for startups, not just their mean.

Tables 1 and 2 display summary statistics for our sample of startups during the first week and the last (173rd week) week of our panel. The mean startup age at the start of our study is 4.14 years, and 37% of startups have raised angel or VC funding. While 16.75% of our startups use A/B testing at some point, only 8% use an A/B testing tool at the start of our panel and 6% at the end. The decrease is primarily due to Optimzely dropping its free plan in December 2017, which led hundreds of Optimzely users to drop the service. While all firms had non-zero page views in the first month of our panel. Consistent with the stylized fact that many startups fail, nearly 28% of startups have persistent zero page view weeks by the end of our observation window.

Results

The effect of A/B testing on startup performance

We begin our analysis by examining whether the adoption of A/B testing is related to the growth in a startup's number of page views. These results are presented in Table

3.

Model 1 in Table 3 estimates the raw correlation between the lagged adoption of A/B testings and weekly page views. This model does not account for any possible confounding that may be due to non-random selection into A/B testing or omitted variables. This correlation is relatively significant, and we see that firms that use A/B testing have 280% more page views than those that do not. In model 2, we account for firm-level heterogeneity by including firm fixed effects. This regression shows that the majority of this original estimate is due to selection. After controlling for time-invariant differences between firms, the estimated effect of A/B testing drops dramatically to 46%. Thus, a simple comparison between firms using A/B testing tools and those that do not would yield an overly optimistic view about the value of A/B testing for firm outcomes.

Model 3 controls for time-varying heterogeneity by including a control for the size of each firm's technology stack. If the adoption of A/B testing is correlated with the adoption and use of other technologies (e.g., payment processing tools), then the estimate might reflect the impact of different technologies and not of A/B testing or major technological pivots that lead to better performance. By controlling for the time-varying size of a startup's technology stack, we can isolate the adoption of A/B tools from other technologies. After controlling for technology stack size, the estimated effect of A/B testing is now just under 11.8%. Firms that are adopting more technology over time are also adopting A/B testing tools.

Model 4 further accounts for time-varying heterogeneity by including fixed effects for each firm's growth trajectories. The impact of these heterogeneous firm slopes is more marginal for our coefficient of interest. The estimated effect of A/B testing drops to 8.5%. However, the difference between the estimates in models 3 and 4 is not itself statistically significant. It appears that the control for technology stack size accounts for much of the relevant time-varying heterogeneity related to the adoption of A/B testing technology.

Overall, models 1–4 show that firms vary in whether they adopt A/B testing tech-

nology. However, when we account for these selection effects, we find that the residual impact of adopting A/B testing for the marginal firm is roughly 10%. This table provides preliminary evidence that startups that adopt A/B testing do indeed achieve a significant improvement in their performance as measured by page views.

Figure 2 shows the effect of A/B testing graphically. To do so, we restrict our sample to the 4,645 firms that adopt an A/B testing tool within our observation window. We then estimate the effect of A/B testing for each quarter after *and before* the firm adopts. Doing so allows us to explore how quickly a firm benefits from testing and allows us to test for the presence of pre-trends before adoption. Specifically, we collapse our data down to the quarterly level and then fit a model with firm fixed effects, quarter fixed effects, and a technology stack control. We then include dummies for whether the observation is N quarters before or after the firm adopts. Quarter 0 is the quarter *before* the firm adopts and serves as a our excluded baseline. Figure 2 shows the estimates and confidence intervals from this model. In the year before adoption, the effect of A/B testing is negative and is not statistically different from zero. There does not appear to be any upward trend. Post-adoption, it takes roughly one year for effects to materialize. Consistent with our fixed effect estimates, the effect on page view growth after one year is roughly 8%. In Appendix Table A1 we show that the fixed effects specification in Table 3 holds when run on this event study sample.

Does A/B testing lead to new product introductions?

A further prediction from our theory is that reducing the cost of testing ideas should also increase the rate at which startups generate new ideas to test—i.e., product innovating. Moreover, with an increase in idea generation, we should also observe startups that adopt A/B testing having more "tail" outcomes—i.e., failing and scaling—at higher rates. Table 4 tests these arguments.

Consistent with our model of A/B testing, in model 1, we find that using A/B testing is related to a new 0.067 product launches. In model 2, we re-estimate our product launch regression with a truncated distribution (at 10 introductions) to ensure

our results are not driven by the extreme tail of this distribution. Together, our findings suggest that A/B testing increases the number of product launches by 9% to 18%.

Finally, in models 3 and 4, we estimate the impact of A/B testing on the likelihood that startups have extreme outcomes. We find that the adoption of A/B testing increases the likelihood of startups achieving both left- and right-tail outcomes—e.g., zero page views and 100k+ page views—after they adopt A/B testing. This finding is consistent with our prediction that A/B testing helps startups recognize what works and also that nothing works, thereby increasing the likelihood that they fail and scale faster. Specifically, we find that A/B testing increases the likelihood that a firm receives no page views by 0.7 percentage points or hitting 100,000 page views (the 95th percentile of views) by 1.4 percentage points.

Who benefits from A/B testing?

In Table 5, we test our prediction that the effect of A/B testing should depend on the degree to which the startup is more or less likely to have product-market fit. In model 1, we show that consistent with the arguments above, older startups are more likely to benefit from A/B testing relative to younger ones.[7] In Figure 3, we show how the estimated treatment effect varies by startup age. Since experience also impacts the estimated effect of A/B testing, we show effects while assuming that the founding team has no prior experience. We find that for newer startups—the youngest in our sample were founded two years ago—appear to actually do worse when using A/B testing, though the 95% confidence interval overlaps with zero. For firms founded more than roughly five years ago, the effects of A/B testing are positive and significant. Similarly, model 2 shows that the impact of A/B testing on the number of product launches holds primarily for the older startups in our sample.

However, per our theory, we find that A/B testing accelerates development toward rapid growth or pivoting for older and younger startups, respectively. In model 3, we

[7]The number of observations drops to 14,238 startups in Table 4 because we have founder work experience infromation only for this smaller set of firms. Our findings for age are the same when we use the full sample of 35,918 startups.

find that younger firms are more likely to "fail fast," having an increased likelihood of persistent zero page views after adopting A/B testing. In contrast, more mature startups are more likely to scale after adopting A/B testing—having an increased likelihood of 100k+ page views.

Finally, in Table 5, we also test our prediction that the effect of A/B testing will depend on whether founders of the startup have relevant experience that they can use to design and implement better experiments. In model 1, we find some evidence consistent with the idea that founders with relevant experience—having previously founded a startup–derive greater returns to experimentation. Figure 4 shows how the estimated effect of A/B testing varies with prior experience for a startup that is four years old. We find that teams with at least one founder who has worked at another startup benefit from A/B testing. However, unlike the effects of age, we do not find that startups with more experienced founders introduce significantly more new product features. Nor do we find that such startups experience tail outcomes such as an increased likelihood of persistent zero page views or 100K+ page views. On the whole, there is only modest evidence that prior experience leads to better implementation of an experimental strategy.

Robustness tests

In this section, we summarize several additional tests we conducted to account for alternative explanations. In particular, because we leverage an observational research design without experimental variation in who adopts A/B testing, we must account for several important factors that may bias our estimates upward.

To summarize, in our estimations described above, we account for several relevant sources of endogeneity in our baseline models. First, our results account for reverse causality by using panel data that tests the impact on growth after adoption of A/B testing. Second, we consider unobserved selection into the adoption of A/B testing using a rich set of fixed effects. Our models include firm fixed effects that account for systematic variation in the levels of performance across startups that may be due to

differences in founding teams, initial funding, industry, and location. Our technology stack control rules accounts for time-varying heterogeneity in the other technology tools a firm chooses to adopt. Moreover, we use heterogeneous firm slopes in Table 3, model 4 to account for the possibility that some startups may have different growth trajectories than others. These varied trajectories may be due to differences in initial investments or strategy (e.g., using digital advertising to grow fast). Even with these controls, we find strong evidence that the adoption of A/B testing is strongly related to performance. Finally, Figure 2 shows that before adoption of A/B testing, the effect is flat and that it is only after adoption that we see any impact of A/B testing on page view growth.

To further test our modeling strategy, we conduct a placebo test in Appendix Section A2 where we substitute our A/B testing adoption measure with the adoption of a cloud font library. While new fonts sourced from the cloud might improve startup growth by improving a page's design, we expect the average effect to be much smaller and close to zero. We find no evidence that adopting new fonts has a significant effect on performance. Our modeling strategy does not automatically generate positive effects of technology adoption on performance.

We also address the concern that our results might be driven by our choice of growth metrics in Appendix Section A3. The SimilarWeb data also provide weekly information about other performance metrics for startups, including a website's time on site and pages per visit. The Crunchbase data also include time-varying data on the total amount of VC and angel funding a startup has raised. We find evidence consistent with our main models that shows that the adoption of A/B testing increases time on site, the number of pages visited by a user, and the amount of funding raised post-adoption.

Finally, we conduct two additional robustness tests. In Appendix Section A4, we show that our findings hold when we restrict our sample to the 13,186 startups that had already raised funding by March 2015. Our data tends to have stronger coverage of these startups, and it is heartening to see our findings hold when there are fewer

measurement error and missing data concerns. Finally, in Appendix Section A5, we test if our results are driven by the specific A/B testing tools we selected. To do so, we replicate Table 3 but include not just A/B-focused tools but any web technology that integrates or enables A/B testing technology. Using this more expansive measure yields similar results.

Together, these results offer strong support to our claim that the adoption of A/B testing is related to improved startup performance.

Discussion and Conclusion

What is the right entrepreneurial strategy for startups? Recent work in strategic management identifies experimentation as the best framework for decision making in young businesses (Levinthal, 2017; Camuffo et al., 2019; Gans, Stern and Wu, 2019). 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 startup performance. We build a unique dataset of over 35,000 global startups, their adoption of A/B testing, and measures of weekly performance.

We find that while only 7.6% of startups used A/B testing technology at the beginning of our panel but that proportion more than doubled over the subsequent eight years. A/B testing is associated with a 10% increase in page views and is positively correlated with increased product introductions. Interestingly, as our theoretical arguments predicted, we find that A/B testing is also associated with startups scaling and failing faster. This finding supports the idea that A/B testing enables more decisive actions in both the right and left tails of startup performance. However, not all startups benefit equally from A/B testing. We find partial evidence that our effects are more pronounced for startups with more experienced founders. Moreover, it is the younger startups using A/B testing that fail, whereas their older counterparts leverage this technology to scale.

Our article informs two research agendas at the intersection of strategy and en-

trepreneurship. First, while our field has generated many insights about strategy in large organizations, it is only recently that we have sought to clarify entrepreneurial strategy. Our findings provide empirical evidence that an experimental approach to strategy, as suggested by Levinthal (2017); Gans, Stern and Wu (2019); Camuffo et al. (2019), is positively associated with better performance. We delve into the underlying mechanisms that support this relationship. We argue and demonstrate that a decline in the cost of testing ideas sparks idea generation and more decisive action in the spirit of the three-part experimentation framework suggested by Knudsen and Levinthal (2007); March (1991). We offer a novel insight that highlights an important distinction between running *an* experiment versus a strategy based on *experimentation*. Running a single experiment will most likely to lead to null or negative results, since most ideas fail (Kohavi and Longbotham, 2017; Kohavi and Thomke, 2017). A strategy based on repeated experimentation will over time reveal unexpected high-quality ideas that will improve performance.

Another contribution of our work is to the emerging literature on data-driven decision-making and the broader digitization of the global economy (Brynjolfsson, Hitt and Kim, 2011; Brynjolfsson and McElheran, 2016). This literature has argued that the vast amount of transaction data collected by firms allows them to do unprecedented analysis of consumer data to inform their strategies. We demonstrate that A/B testing enables firms to do more than simply analyze the past. By generating, testing, and implementing *new* ideas, firms can use the data created through digital experimentation to design the future.

Our approach is not without limitations. While we build the first large-panel dataset on startup experimentation, we recognize that A/B testing is not randomly assigned to firms. This selection challenge could bias our estimates upward, though we take care to control for important observed and unobserved factors that might drive A/B testing adoption and performance. In our most demanding specifications, we include controls for the adoption of other technologies, a rich suite of fixed effects, and also heterogeneous firm-level growth trajectories. The dynamic panel structure of our data, coupled

with these controls, allows us to rule out various kinds of selection bias, reverse causality, and omitted variable bias. Our results, demonstrated on multiple performance metrics, are in line with previous studies on the effect of data-driven decision making on firm performance (Brynjolfsson, Hitt and Kim, 2011).

Further, we do not observe the A/B tests that startups actually run, which prevents a deeper understanding of the mechanisms connecting firm performance and experimentation. Moreover, our sample is composed of digital firms, which can experiment at low cost. The cost of experimentation is still high in many other industries, though innovations like 3-D printing may be changing this dynamic. While we do consider founder experience, we are unable to evaluate how the adoption of A/B testing influences intra-firm dynamics. We conjecture that these tools will 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.

Moving forward, the continued decline in the cost of running digital experiments will raise 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, while relatively few firms currently do digital experiments, will widespread adoption alter the benefits for an individual organization? Finally, how will experimentation across the economy change the types of innovations developed and how they are distributed? Addressing these questions, among others, will guide future research and practice.

References

Agarwal, Rajshree, Raj Echambadi, April M Franco and Mitrabarun B Sarkar. 2004. "Knowledge transfer through inheritance: Spin-out generation, development, and survival." *Academy of Management journal* 47(4):501–522.

Aral, Sinan and Dylan Walker. 2011. "Creating social contagion through viral product design: A randomized trial of peer influence in networks." *Management science* 57(9):1623–1639.

Aral, Sinan and Dylan Walker. 2014. "Tie strength, embeddedness, and social influence: A large-scale networked experiment." *Management Science* 60(6):1352–1370.

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

Arrow, Kenneth J. 1962. "The Economic Implications of Learning by Doing." *The Review of Economic Studies* 29(3):155–173.

Azevedo, Eduardo M, Deng Alex, Jose Montiel Olea, Justin M Rao and E Glen Weyl. 2018. "A/B Testing with Fat Tails.".

Bapna, Ravi, Jui Ramaprasad, Galit Shmueli and Akhmed Umyarov. 2016. "One-way mirrors in online dating: A randomized field experiment." *Management Science* 62(11):3100–3122.

Bhide, Amar. 1986. "Hustle as strategy." *Harvard Business Review* 64(5):59–65.

Bhidé, Amar V. 2003. *The origin and evolution of new businesses*. Oxford University Press.

Blank, Steve. 2013. *The four steps to the epiphany: successful strategies for products that win*. BookBaby.

Brynjolfsson, Erik and Andrew McAfee. 2012. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Brynjolfsson and McAfee.

Brynjolfsson, Erik and Kristina McElheran. 2016. "The rapid adoption of data-driven decision-making." *American Economic Review* 106(5):133–39.

Brynjolfsson, Erik and Lorin M Hitt. 2003. "Computing productivity: Firm-level evidence." *Review of economics and statistics* 85(4):793–808.

Brynjolfsson, Erik, Lorin M Hitt and Heekyung Hellen Kim. 2011. "Strength in numbers: How does data-driven decisionmaking affect firm performance?" *Available at SSRN 1819486* .

Camerer, Colin and Dan Lovallo. 1999. "Overconfidence and excess entry: An experimental approach." *American economic review* 89(1):306–318.

Camuffo, Arnaldo, Alessandro Cordova, Alfonso Gambardella and Chiara Spina. 2019. "A scientific approach to entrepreneurial decision-making: Evidence from a randomized control trial." *Forthcoming in Management Science* .

Chatterji, Aaron K and Kira Fabrizio. 2012. "How do product users influence corporate invention?" *Organization Science* 23(4):971–987.

Cohen, Wesley M and Daniel A Levinthal. 1994. "Fortune favors the prepared firm." *Management Science* 40(2):227–251.

David, Paul A et al. 1975. *Technical choice innovation and economic growth: essays on American and British experience in the nineteenth century*. Cambridge University Press.

Dubé, Jean-Pierre, Zheng Fang, Nathan Fong and Xueming Luo. 2017. "Competitive price targeting with smartphone coupons." *Marketing Science* 36(6):944–975.

Fabijan, Aleksander, Pavel Dmitriev, Helena Holmström Olsson and Jan Bosch. 2017. The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale. In *Proceedings of the 39th International Conference on Software Engineering*. IEEE Press pp. 770–780.

Gans, Joshua S, Scott Stern and Jane Wu. 2019. "Foundations of entrepreneurial strategy." *Strategic Management Journal* 40(5):736–756.

Ghemawat, Pankaj. 1991. *Commitment*. Simon and Schuster.

Ghemawat, Pankaj and Patricio Del Sol. 1998. "Commitment versus flexibility?" *California Management Review* 40(4):26–42.

Girotra, Karan, Christian Terwiesch and Karl T Ulrich. 2010. "Idea generation and the quality of the best idea." *Management science* 56(4):591–605.

Gomez-Uribe, Carlos A and Neil Hunt. 2016. "The netflix recommender system: Algorithms, business value, and innovation." *ACM Transactions on Management Information Systems (TMIS)* 6(4):13.

Gompers, Paul, Josh Lerner and David Scharfstein. 2005. "Entrepreneurial spawning: Public corporations and the genesis of new ventures, 1986 to 1999." *The journal of Finance* 60(2):577–614.

Hannan, Michael. 1984. "Structural Inertia and Organizational Change." *American Sociological Review* 49(2):149–164.

Hendel, Igal and Yossi Spiegel. 2014. "Small steps for workers, a giant leap for productivity." *American Economic Journal: Applied Economics* 6(1):73–90.

Kaplan, Steven N and Josh Lerner. 2016. Venture capital data: Opportunities and challenges. Technical report National Bureau of Economic Research.

Klepper, Steven and Sally Sleeper. 2005. "Entry by spinoffs." *Management science* 51(8):1291–1306.

Knudsen, Thorbjørn and Daniel A Levinthal. 2007. "Two faces of search: Alternative generation and alternative evaluation." *Organization Science* 18(1):39–54.

Kohavi, Ron, Randal M Henne and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 959–967.

Kohavi, Ron and Roger Longbotham. 2017. "Online controlled experiments and a/b testing." *Encyclopedia of machine learning and data mining* pp. 922–929.

Kohavi, Ron and Stefan Thomke. 2017. "The surprising power of online experiments." *Harvard Business Review* 95(5):74–+.

Kohavi, Ronny, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres and Tamir Melamed. 2009. "Online experimentation at Microsoft." *Data Mining Case Studies* 11.

Kumar, Anuj and Yinliang Tan. 2015. "The demand effects of joint product advertising in online videos." *Management Science* 61(8):1921–1937.

Lawrence, Alastair, James Ryans, Estelle Sun and Nikolay Laptev. 2018. "Earnings announcement promotions: A Yahoo Finance field experiment." *Journal of Accounting and Economics* 66(2-3):399–414.

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

Levitt, Steven D, John A List and Chad Syverson. 2013. "Toward an understanding of learning by doing: Evidence from an automobile assembly plant." *Journal of Political Economy* 121(4):643–681.

March, James G. 1991. "Exploration and exploitation in organizational learning." *Organization science* 2(1):71–87.

McGrath, Rita Gunther. 1999. "Falling forward: Real options reasoning and entrepreneurial failure." *Academy of Management review* 24(1):13–30.

McGrath, Rita Gunther and IC MacMillan. 2000. "The entrepreneurial mindset: Strategies for continuously creating opportunity in an age of uncertainty (Vol. 284).".

McMullen, Jeffery S and Dean A Shepherd. 2006. "Entrepreneurial action and the role of uncertainty in the theory of the entrepreneur." *Academy of Management Review* 31(1):132–152.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of general psychology* 2(2):175–220.

Nordhaus, William D. 2007. "Two centuries of productivity growth in computing." *The Journal of Economic History* 67(1):128–159.

Ries, Eric. 2011. *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Books.

Sahni, Navdeep S, Dan Zou and Pradeep K Chintagunta. 2016. "Do targeted discount offers serve as advertising? Evidence from 70 field experiments." *Management Science* 63(8):2688–2705.

Sarasvathy, Saras D. 2001. "Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency." *Academy of management Review* 26(2):243–263.

Shane, Scott. 2000. "Prior knowledge and the discovery of entrepreneurial opportunities." *Organization science* 11(4):448–469.

Simon, Herbert A. 1959. "Theories of Decision-Making in Economics and Behavioral Science." *American Economic Review* 49(3):253–283.

Siroker, Dan, Pete Koomen, Elliot Kim and Eric Siroker. 2014. "Systems and methods for website optimization.". US Patent 8,839,093.

Sitkin, Sim B. 1992. "Learning through failure: The strategy of small losses." *Research in organizational behavior* 14:231–266.

Sørensen, Jesper B and Toby E Stuart. 2000. "Aging, obsolescence, and organizational innovation." *Administrative science quarterly* 45(1):81–112.

Thomke, Stefan. 2001. "Enlightened experimentation: The new imperative for innovation." *Harvard Business Review* 79(2):66–75.

Urban, Steve, Rangarajan Sreenivasan and Vineet Kannan. 2016. *It's All A/Bout Testing: The Netflix Experimentation Platform.*
URL: *https://medium.com/netflix-techblog/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15*

Van den Steen, Eric. 2016. "A formal theory of strategy." *Management Science* 63(8):2616–2636.

Xu, Ya. 2015. *Why Experimentation is so Important for LinkedIn.*
URL: *https://engineering.linkedin.com/ab-testing/why-experimentation-so-important-linkedin*

Xu, Ya, Nanyu Chen, Addrian Fernandez, Omar Sinno and Anmol Bhasin. 2015. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM pp. 2227–2236.

Table 1: Summary statistics at Week 1

	Mean	Median	SD	Min	Max	N
Age	4.14	4.00	1.64	2	7	35,918
Funded from start	0.37	0.00	0.48	0	1	35,918
Using A/B tool?	0.08	0.00	0.27	0	1	35,918
Technology Stack	39.09	37.00	17.15	1	80	35,918
Log(Pageviews + 1)	7.40	7.35	2.57	0	20	35,918
Number of product launches	0.00	0.00	0.05	0	1	13,186
0 pageview week	0.04	0.00	0.20	0	1	35,918
100K+ pageview week	0.05	0.00	0.22	0	1	35,918

Table 2: Summary statistics at Week 173

	Mean	Median	SD	Min	Max	N
Age	7.14	7.00	1.64	5	10	35,918
Funded from start	0.37	0.00	0.48	0	1	35,918
Using A/B tool?	0.06	0.00	0.24	0	1	35,918
Technology stack	41.96	38.00	30.83	1	477	35,918
Log(Pageviews + 1)	5.65	6.40	4.04	0	20	35,918
Number of product launches	0.36	0.00	2.44	0	133	13,186
0 pageview week	0.28	0.00	0.45	0	1	35,918
100K+ pageview week	0.05	0.00	0.23	0	1	35,918

Table 3: A/B testing and Startup Growth

	Log(Pageviews + 1)			
	(1)	(2)	(3)	(4)
Using A/B tool?	2.802***	0.464***	0.118***	0.085***
	(0.047)	(0.028)	(0.023)	(0.019)
Observations	6,213,814	6,213,814	6,213,814	6,213,814
Number of startups	35,918	35,918	35,918	35,918
Week FE	Y	Y	Y	Y
Firm FE		Y	Y	Y
Technology Stack Control			Y	Y
Growth FE				Y

Standard errors clustered at the firm level in parentheses.

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

Table 4: A/B testing and product introductions

	(1) # of product launches	(2) # of product launches (truncated)	(3) 0 pageviews	(4) 100K+ pageviews
Using A/B tool?	0.067**	0.033*	0.007**	0.014***
	(0.026)	(0.014)	(0.003)	(0.002)
Observations	2,281,178	2,281,178	6,213,814	6,213,814
Number of startups	13,186	13,186	35,918	35,918
Week FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y

Standard errors clustered at the firm level in parentheses.

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

Table 5: The impact of A/B testing varies with startup age and founder experience

	(1) Log of Pageviews + 1	(2) # of product launches	(3) 0 pageviews	(4) 100K+ pageviews
Using A/B tool?	-0.171*	-0.300**	0.072***	-0.000
	(0.077)	(0.101)	(0.008)	(0.010)
Using A/B Technology X Age	0.047***	0.074***	-0.011***	0.003*
	(0.012)	(0.020)	(0.001)	(0.002)
Using A/B Technology X Experience	0.098*	0.015	-0.001	-0.001
	(0.048)	(0.050)	(0.005)	(0.006)
Observations	2,463,174	1,281,757	2,463,174	2,463,174
Number of startups	14,238	7,409	14,238	14,238
Week FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y

Standard errors in parentheses

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

Figure 1: Example of the BuiltWith Data

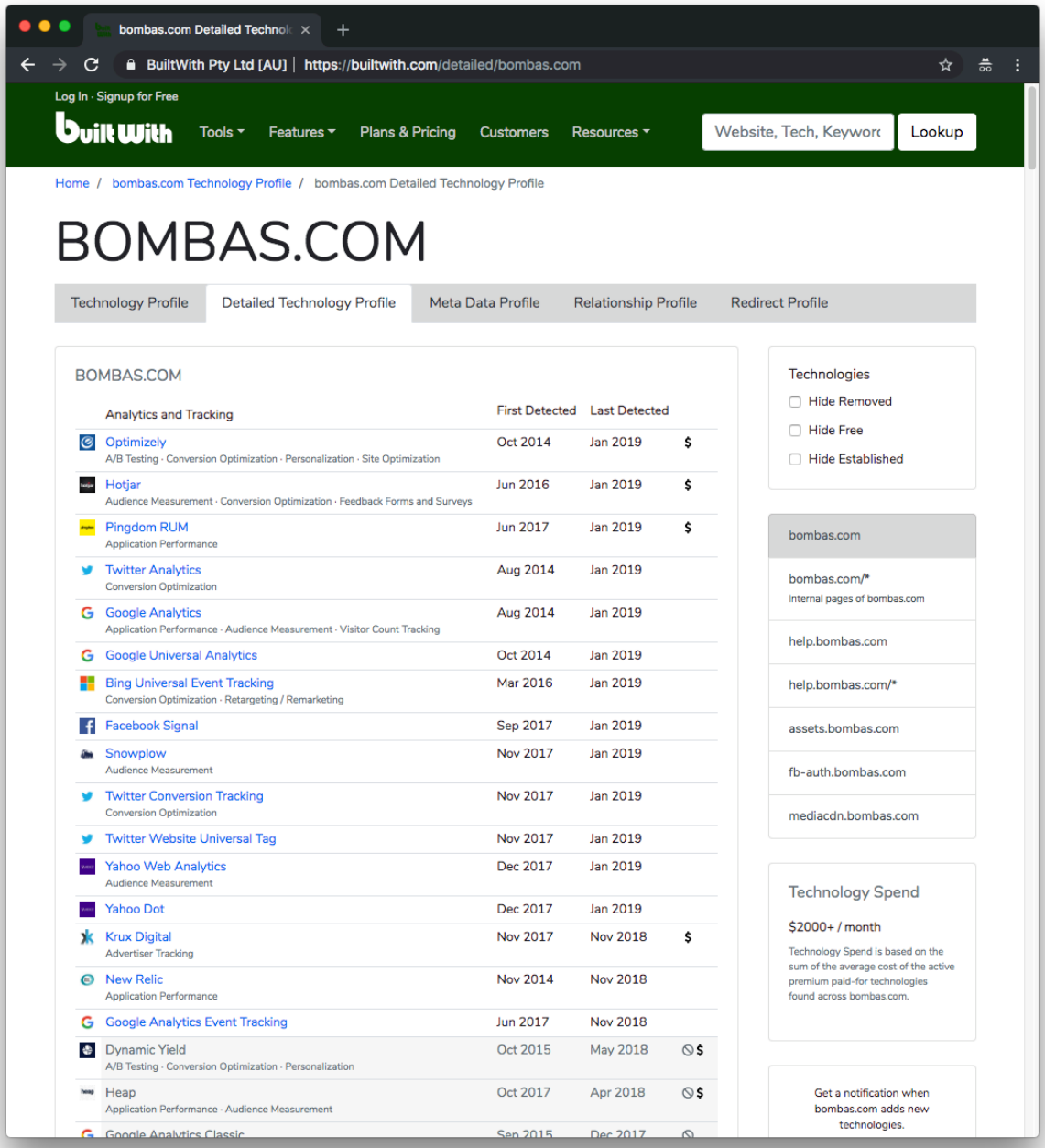


Figure 2: Plot of estimates and confidence intervals from the event study specification described in the paper. Quarter 1 is the first quarter a firm is observed using an A/B testing tool. We use the quarter just before adoption (quarter=0) as the excluded baseline.

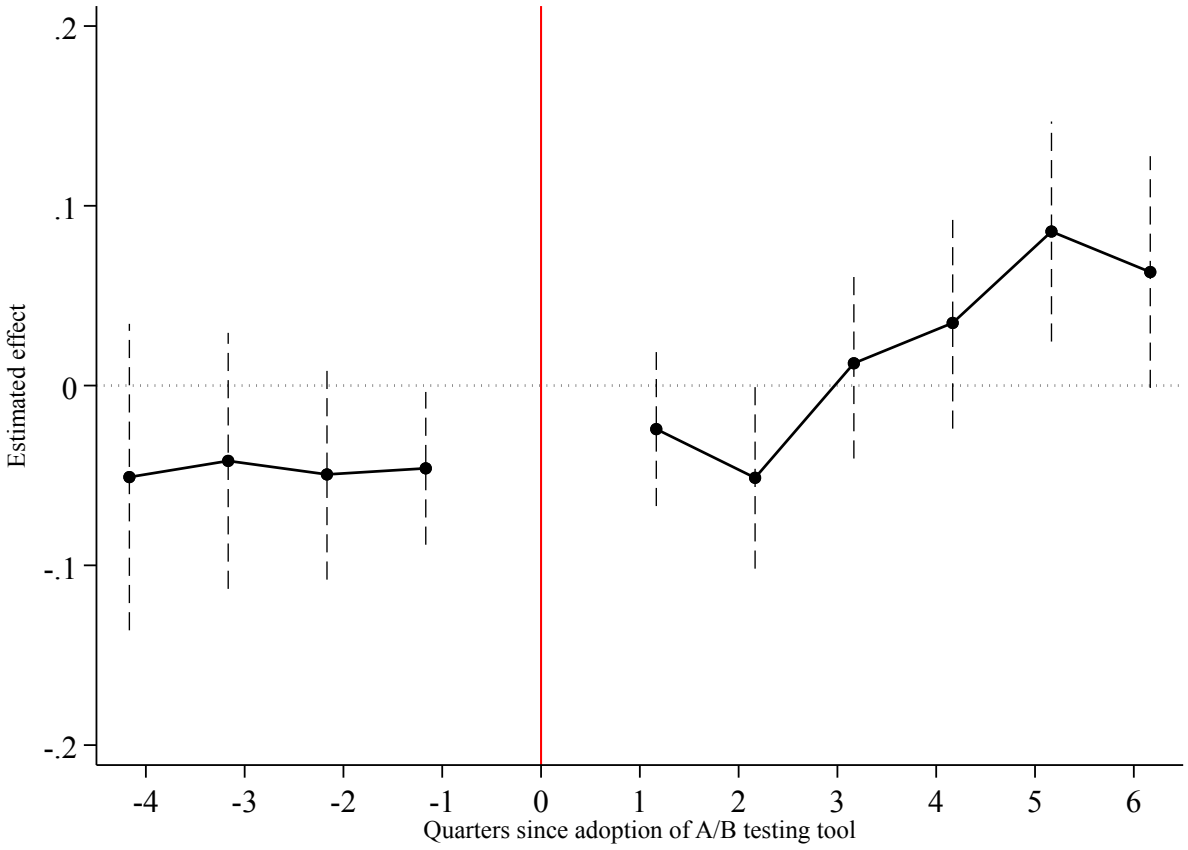


Figure 3: Estimated effects by age for a startup with no prior experience

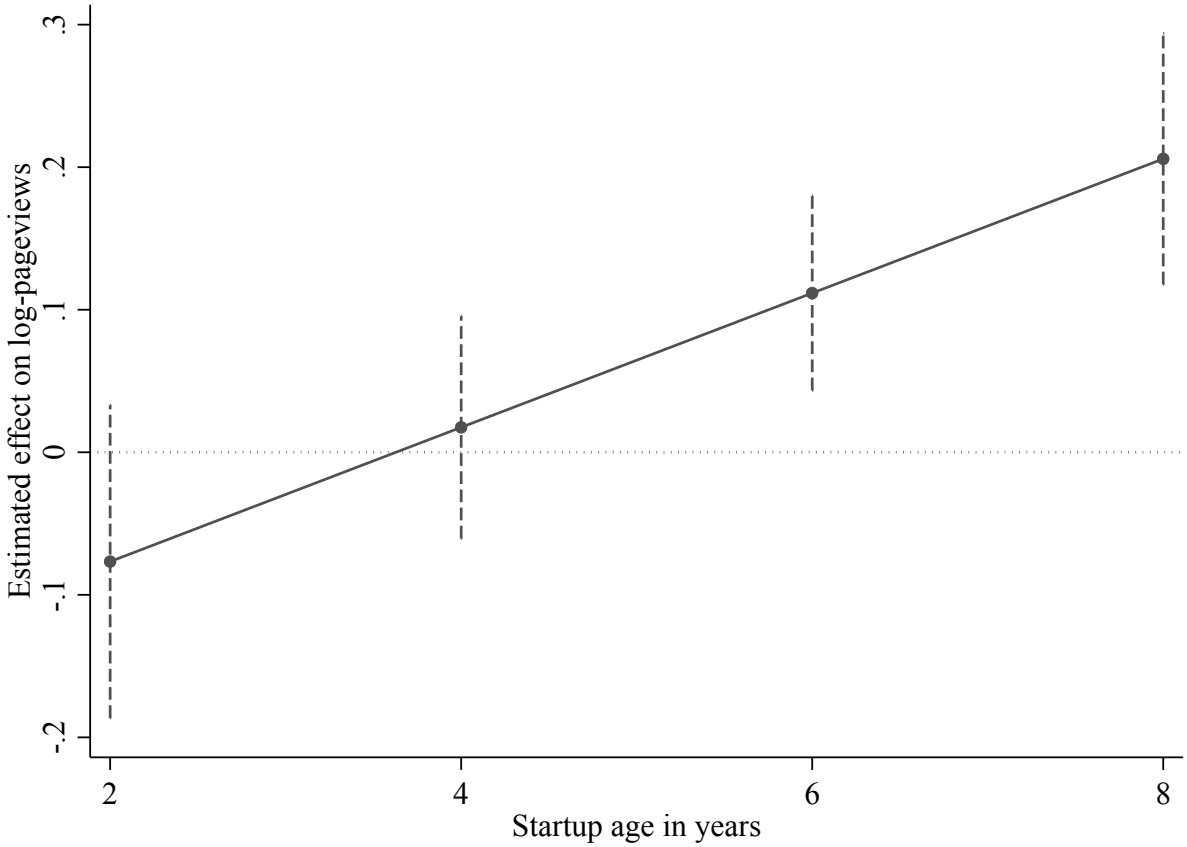
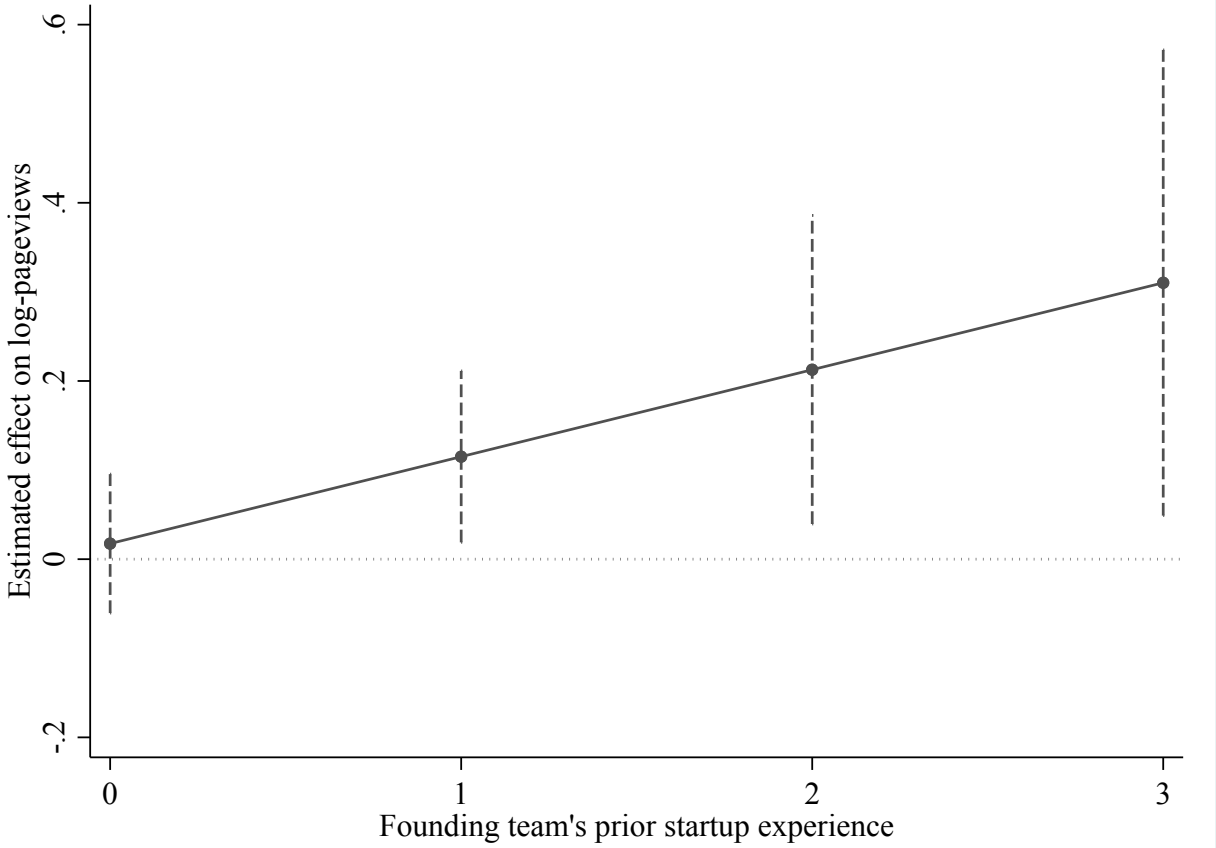


Figure 4: Estimated effects by experience for a four year old startup.



A1 Quarterly event study models

As described in the the paper, we graphically test for pre-trends by focusing only on firms that adopt A/B testing tools once. Table A1 focuses on these 4,645 firms and replicate our primary growth models in Table 3. Our findings hold when we identify the week and technology stack controls only on adopting firms.

A2 Placebo test using font libraries

In Table A2 we run a placebo test. Specifically, we show that adopting a new cloud-based font library has no impact on startup growth once we account for firm fixed effects, week fixed effects and our technology stack control. We have no reason to believe that adding a could font library should have a causal impact on growth. While faster growing or larger firms might be more likely to use a cloud font provider, the adoption of this tool should, at best, have a minimal effect on growth. Specifically, we test if adopting any of the following tools has an effect: *google font api, font awesome, fonts.com, myfonts, adobe creative cloud webfonts, webtype, font awesome, mozilla fonts, mozilla fonts, fonts.com, adobe edge web fonts, and fontplus*. Model 1 shows that larger startups are more much more likely to use cloud font libraries. Model 2 shows that including firm fixed effects reduces the impact, but a positive and significant coefficient remains. Model 3 includes our technology stack control (excluding cloud font tools). We find a well estimated zero indicating that our preferred specification does not necessarily return a positive effects, even when a tool should have no impact on growth. In Model 4 we include growth fixed effects and find, if anything, a small negative effect.

A3 Alternative performance metrics

Table A3 tests the impact of A/B testing on alternative growth and performance measures. Model 1 tests the impact on the bounce rate. The bounce rate is the percent of visitors who navigate away from a website after visiting only one page. Model 2 tests the impact on the logged average number of pages a browser visits. Model 3 looks at the logged of average visit duration in seconds. Finally, Model 4 looks at the logged total VC and angel funding raised by the startup. While we find no impact on the bounce rate, we find A/B testing increases the pages per visit, the amount of time spent on the website, and the amount of funding raised by the startup. The effect sizes range from 1.4% to 5.5%.

A4 Funded startups only

Table A4 replicates Table 3 using only the startups that had already raised funding at the start of our panel. Our findings hold on this this smaller sample.

A5 Impact of A/B enabled tools

In Table A5 we show our findings hold when using a more expansive definition of A/B testing tools. In the body of the paper, we focus on tools that only and explicitly focus on A/B testing. Here, we include both these core A/B testing tools and analytics tools that enable and/or integrate with A/B testing technologies. Specifically, we look at the following tools: *ab tasty, adobe target standard, avenseo, beampulse, bunting, changeagain, conductrics, convert, devatics, dynamic yield, experiment.ly, google content experiments, google optimize 360, google website optimizer, iterable, kaizen platform, kameleoon, kissmetrics, leanplum, marketizator, maxymiser, maxymiser, maxymizely, mixpanel, monetate,*

monetate, myna, omniture adobe test and target, optimization robot, optimizely, optimost, qubit deliver, roistat, sentient ascend, shopalize, sigopt, sitegainer, site-spect, split optimizer, stetic, visual website optimizer, visual website optimizer, and *zarget.*

Table A1: Fixed effects results for the event study sample used to generate Figure 2. The event study sample is at the quarterly level for firms that adopt A/B testing tools only once.

	Log(Pageviews + 1)			
	(1)	(2)	(3)	(4)
Using A/B tool?	0.851***	0.478***	0.049*	0.085***
	(0.056)	(0.028)	(0.023)	(0.020)
Observations	65,030	65,030	65,030	65,030
Number of startups	4,645	4,645	4,645	4,645
Quarter FE	Y	Y	Y	Y
Firm FE		Y	Y	Y
Technology Stack Control			Y	Y
Growth FE				Y

Standard errors clustered at the firm level in parentheses.

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

Table A2: Placebo test checking if adopting a new cloud-based font library impacts startup growth.

	Log(Pageviews + 1)			
	(1)	(2)	(3)	(4)
Using Cloud Font Library?	1.245***	0.888***	0.000	-0.045**
	(0.032)	(0.018)	(0.016)	(0.015)
Observations	6,213,814	6,213,814	6,213,814	6,213,814
Number of startups	35,918	35,918	35,918	35,918
Week FE	Y	Y	Y	Y
Firm FE		Y	Y	Y
Tech. Stack Control			Y	Y
Growth FE				Y

Standard errors clustered at the firm level in parentheses.

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

Table A3: A/B testing improves alternative growth and performance metrics including the number of pages visited, time on website, and the amount of VC/Angel funding raised. It has no measurable effect on the bounce rate.

	(1) Bounce Rate	(2) Log Pages per Visit	(3) Log Visit Duration	(4) Log Funding Raised
Using A/B tool?	-0.003	0.014**	0.042**	0.055*
	(0.002)	(0.005)	(0.013)	(0.027)
Observations	4,601,773	4,601,647	4,601,589	6,213,814
Number of startups	35,918	35,918	35,918	35,918
Week FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y

Standard errors clustered at the firm level in parentheses.

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

Table A4: The impact of A/B testing holds when we focus only on startups that had already raised funding at the start of our panel.

	Log(Pageviews + 1)			
	(1)	(2)	(3)	(4)
Using A/B tool?	2.858***	0.485***	0.088**	0.068**
	(0.061)	(0.038)	(0.031)	(0.026)
Observations	2,281,178	2,281,178	2,281,178	2,281,178
Number of startups	13,186	13,186	13,186	13,186
Week FE	Y	Y	Y	Y
Firm FE		Y	Y	Y
Technology Stack Control			Y	Y
Growth FE				Y

Standard errors clustered at the firm level in parentheses.

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

Table A5: Using a more expansive definition of A/B testing yields similar results as our more focused definition.

	Log(Pageviews + 1)			
	(1)	(2)	(3)	(4)
Using Technology with A/B Feature?	2.534***	0.620***	0.162***	0.141***
	(0.039)	(0.024)	(0.020)	(0.017)
Observations	6,213,814	6,213,814	6,213,813	6,213,813
Number of startups	35,918	35,918	35,918	35,918
Week FE	Y	Y	Y	Y
Firm FE		Y	Y	Y
Technology Stack Control			Y	Y
Growth FE				Y

Standard errors clustered at the firm level in parentheses.

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