Digital 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

February 11, 2021

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 on how digital experimentation affects a large sample of high-technology startups using data that tracks their growth, technology use, and product launches. Despite its prominence in the business press, relatively few firms have adopted A/B testing tools. However, among those that do, we find increased performance on several critical dimensions, including page views, code changes, and new product features. These results are robust to our use of instrumental variables and synthetic control models, where we leverage Google's launch of a new A/B testing tool. Our results also show that the value of A/B testing increases over time and is positively related to tail outcomes, with adopting firms both scaling and failing faster. Our results inform the emerging literature on entrepreneurial strategy and illustrate the growing importance of digitization and data-driven decision-making.

*Authors names are in reverse alphabetical order. All authors contributed equally to this project. We thank seminar participants at the Harvard Business School, the Conference on Digital Experimentation, Duke University, University of Maryland, Binghamton University, University of Minnesota, NYU, The Utah Strategy Conference, 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 organizational learning approach to entrepreneurial strategy centered on experimentation. He envisions 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 to de-bias learning. 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), 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 cost of experimentation 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 and learn from them. Accelerators, venture capital firms, and leading entrepreneurs advocate that startups should A/B test nearly everything they do and incorporate what they learn into their strategies.

However, while A/B testing has undoubtedly reduced the cost of evaluating competing ideas, it is an open question whether it facilitates organizational learning—a

more expansive concept. Indeed, prior literature suggests that organizations have long learned from a variety of sources, including other firms (e.g., Mowery, Oxley and Silverman, 1996), outside experts (Jeppesen and Lakhani, 2010), customers (e.g., Chatterji and Fabrizio, 2014; Urban and Von Hippel, 1988; Dahlander and Piezunka, 2014; Cohen, Nelson and Walsh, 2002), suppliers (e.g., Dyer and Hatch, 2004), peers (e.g., Chatterji et al., 2019), and even their own failures (e.g., 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. However, prior scholarship has also highlighted that the learning process is often biased (e.g., Denrell and March, 2001; Denrell, 2003). Firms often generate limited alternatives, rely on ad-hoc processes to evaluate ideas and make decisions based on expediency 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; Camuffo et al., 2019; Gans, Stern and Wu, 2019). For digital experimentation to matter, firms must also generate many alternatives to test *and* let the data drive decision-making (Brynjolfsson and McElheran, 2016).

Moreover, we posit that the impact of experimentation on performance depends on what the firm is testing. There is an ongoing debate among scholars and practitioners about whether A/B tests *merely* drive incremental change or enable more radical shifts in product strategy. At its core, one A/B experiment is incremental—i.e., it tests a narrow hypothesis about the effect of a product change on a handful of metrics such as clicks, revenue, or user retention (Thomke, 2020). While many incremental tests can help a firm "hill climb" with an existing product, it may also blind them to more significant shifts in their industry.

Alternatively, some have suggested that low-cost A/B testing may enable a culture of experimentation that leads to radical innovation. 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 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 adopting A/B testing in just over 35,000 global startups. Our data combines a panel of consistent measures of when firms adopt A/B testing with weekly performance measures between 2015 and 2018. We complement this data with a rich set of information about each startup's technology stack, funding, product characteristics, and website code. We take an abductive approach by presenting results from a wide range of empirical techniques and drawing on insights from industry practitioners on how startups use A/B testing to drive firm performance (Timmermans and Tavory, 2012; King, Goldfarb and Simcoe, 2019).

Our first finding is that there is considerable heterogeneity in the adoption of A/B testing technology. At any given point in time, roughly 8% of startups use an A/B testing technology. However, nearly 18% of startups adopt an A/B testing tool during our four-year panel. Across various demanding fixed-effect specifications, we find that A/B testing is associated with a persistent 5-20% increase in page visits after adoption. Additional tests support the theory that reducing the cost of experimentation A/B testing induces firms to make complementary changes in their strategy. After adopting A/B testing tools, startups are more likely to change their web site's code, more likely to make both incremental and radical code changes, and more likely to introduce new products. Indeed, 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 experimenting is associated with an increased likelihood of tail outcomes—i.e., more zero page-view weeks and more 50k+ page view weeks.

We also study heterogeneity in the magnitude of A/B testing—including factors such as venture funding, age, location, and size. Overall, we find little evidence of significant differences in this technology's benefits across different kinds of startups. The most pronounced difference is among younger and older startups, with younger

startups appearing to benefit more. Similarly, we find some heterogeneity in effect sizes across industry categories, with startups in e-commerce, software, financial services, education, healthcare, and hardware, perhaps benefiting most.

Finally, we use Google's launch of a new A/B testing tool 24 months into our panel as a shock that allows us to test further the causal impact of A/B testing on startup growth trajectories. Using fixed-effects, instrumental variables, and synthetic control models, we find that the adoption of Google's new A/B testing tool substantially increases startup growth rates. Further, we find strong evidence that the impact of A/B testing more than triples after a year of use. This suggests that commitment and learning may play a key role in whether firms reap experimentation's benefits.

Our findings contribute to the literature in several areas. First, we contribute to research 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 build complementary capabilities 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 infrequently test, although the cost of doing so has declined precipitously. Third, we explain how A/B testing connects to the influential literature on organizational learning. Organizations can learn from a variety of sources, including formal experimentation. We argue that formal experimentation using A/B testing can facilitate both incremental and radical innovation, depending on what is being tested. We also 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). Finally, our work also speaks to a substantial literature in marketing that long highlighted the importance of experimental approaches to market testing for improving firm performance (Urban and Katz, 1983; Boulding, Lee and Staelin, 1994).

Theoretical framework

Experimentation as an entrepreneurial strategy

Uncertainty is endemic to the entrepreneurial process (McMullen and Shepherd, 2006). Entrepreneurs must make many decisions, often with uncertain 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; 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 learning and adaptation, 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. Further, 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). Finally, there are many, arguably less expensive, alternative channels for firms to learn from. As mentioned above, 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 above, firms have rarely used formal experiments to inform business decisions.

How A/B tests are used

In the last decade, however, 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). 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., 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.

Although no prior studies have examined the benefits of A/B testing across many firms, various scholarly and practitioner accounts have described the utility of experimentation inside organizations (Kohavi, Henne and Sommerfield, 2007; Kohavi et al., 2009; Kohavi and Longbotham, 2017; Xu et al., 2015). 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 better understand its evolving customer base and use 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 how drivers respond to different incentives when entering a new geographic market.[2]

How A/B testing facilitates organizational learning

Organizational learning requires more than just a reduction in the cost of testing ideas (Fabijan et al., 2017)—which 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 information technology investments yielded returns for firms when they invested in hiring workers with relevant expertise and changing the firms' organizational routines (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:[3]

> *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*

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

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 muting the impact of failed ideas. For example, Emily Dresner, the CTO of Upside Travel notes:[4]

We can ship MVPs and eliminate poor paths — bad messages, bad landing pages, bad flows — without jeopardizing our current progress.

Further, the adoption of A/B testing spurs idea generation and accelerates implementation. Since firms often lack the data to support implementing a significant change, they tend towards inertia and conformity. Allan Willie, a co-founder of business dashboard startup Klipfolio, discusses how A/B testing allowed them to test assumptions about the right pricing model for their startup, rather than rely on the standard templates used by competitors:[5]

Take the time to understand what pricing model works best for you. Most people don't. I'd say that most companies merely copy their competitors, and for lack of a scientific process, just guess. I feel strongly that a business needs to actively test various models to make sure the price it sets is strategic and aligns with its objectives.

We expect A/B testing to facilitate organizational learning, not merely through directly reducing the cost of testing ideas, but also by increasing incentives for idea generation (Levinthal, 2017) and removing barriers to execution (Gans, Stern and Wu, 2019).

How learning via experimentation leads to improved performance

Prior literature links organizational learning to competitive advantage and better performance (March, 1991). Those firms that learn the fastest will be more likely to build

[4]https://engineering.upside.com/upside-engineering-blog-13-testing-culture-d5a1b659665e
[5]https://www.priceintelligently.com/blog/bid/163986/a-complete-guide-to-pricing-strategy

and sustain an edge on their competitors, allowing them to solve complex business challenges and develop new products. However, a contemporary debate among scholars and practitioners questions what firms can learn from A/B tests. A/B tests are often associated with iterative analyses designed to choose the right color on a website or discern the best place to ask for the customer's credit card information. If A/B tests lead to the extension of existing knowledge instead of the creation of new knowledge, how can the adoption of this tool lead to better firm performance?

Prior work has highlighted the complementary role of both exploitation and exploration in supporting organizational learning (March, 1991). A/B testing can rapidly accelerate exploitation, as firms can quickly optimize their best product ideas through relentless continuous improvement at a massive scale. These incremental changes can add up to significant performance improvements (Thomke, 2020). Recent work finds that these kinds of enhancements to existing products account for the vast majority of productivity growth (Garcia-Macia, Hsieh and Klenow, 2019).

Moreover, A/B testing is also useful for breaking down more considerable business challenges into smaller testable hypotheses. By sequentially testing the most critical ideas, product managers can significantly progress toward long-term objectives with uncertain payoffs. For example, engineers at Auth0, recommend an A/B testing strategy where startups "go big, then go small," lest they get stuck in a local maximum:[6]

> *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.*

In addition, A/B testing can make the costs and benefits of radical innovation

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

more transparent, helping managers build a robust quantitative case for a strategic shift. Joel Lewenstein, a product designer at the data collaboration software startup Airtable, describes the value of quantification through A/B testing in helping manage the trade-offs that are a consequence of making big changes.[7]

> *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.*

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 the learning after big strategic decisions, A/B testing can enable radical change. Xu (2015) notes that quantification through A/B helps de-risk 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 in the industry have suggested that the real benefit of experimentation through A/B testing is the ability to conduct clearly defined experiments on radical changes to a product. EJ Lawless, co-founder of Experiment Engine[8], analyzed data on the thousands of A/B tests run on their platform, he finds:[9]

> *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*

[7] 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

[8] Acquired by Optimizely

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

because they are more likely to have a clear hypothesis behind them, while incremental changes are more speculative.

While a single A/B experiment is often assessing an incremental idea, an experiment*ation* strategy using A/B testing can lead to incremental and radical 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. Their study finds that while most tests yield small results, these aggregate into significant gains, and sometimes a single test produces a dramatic improvement (Browne and Jones, 2017). Their findings support our argument that experimentation can enhance firm performance.

How experimentation increases performance variability

When A/B testing leads to radical 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 startup has two natural endpoints: rapidly scaling or failing fast (Yu, 2020). These outcomes are preferable to stagnation or slow growth. Rapid scaling is often necessary for high-technology startups because low entry barriers 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 (Yu, 2020; Camuffo et al., 2019; Arora and Nandkumar, 2011).

A/B testing helps startups recognize which of the natural endpoints they are headed toward. A/B testing increases the number of ideas generated; thus, the variation in their quality also increases (Girotra, Terwiesch and Ulrich, 2010). With an influx of new ideas, good and bad, the challenge of selecting the very best ideas grows as well. Implementing the ideas supported by rigorous experimentation over others will lead to rapid growth. Moreover, as highlighted by Azevedo et al. (2018), Kohavi and Thomke (2017) and (Thomke, 2020), 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 see increases in their average performance. Still, they should also be more likely to experience tail outcomes, scaling, or failing.

The alternative to formal experimentation

Before proceeding to our empirical approach, we consider the appropriate counterfactuals for A/B testing. If startups are not conducting formal experimentation, what other strategies are they taking to organizational learning? Two approaches have been highlighted in prior work. First, an 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 (McGrath, 1999; Nickerson, 1998). The implication is that an overconfident entrepreneur will invest time and effort into implementing strategies that will likely fail (Camuffo et al., 2019). This approach will drive performance differences between firms that experiment and those that do not.

However, a second literature documents that firms have long learned from "uncontrolled experiments" or tweaks to their product development process (David et al., 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, List and Syverson (2013) document a similar phenomenon in an automaker's assembly plant where learning-by-doing (Arrow, 1962) led to productivity gains. In our sample of high-technology startups, firms that are not conducting formal experiments may be tweaking their products informally, which could 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 to tweaking. If, however, tweaking does lead to sustained gains in performance, our A/B testing estimates may have a conservative bias.

Data and Methods

Data sources

To test our predictions, we construct a longitudinal data set comprising 35,262 high-technology startups founded between 2008 and 2013. Our data include information about these startups compiled from four distinct sources. Crunchbase provides us detailed information about each startup's product, funding status, age, location, and team size. We complement the Crunchbase information with weekly measures of page views/visits for each startup from SimilarWeb and use BuiltWith to gather information on the technologies the startups use to build their product, most notably whether and when they use A/B testing tools. Finally, for just under a quarter of our startups, we can collect data on the startup's homepage code over time from the Internet Archive to measure the degree and nature of change associated with adopting A/B testing. 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 industry users' investment/acquisition research. 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, most firms in its sample are startups. The quality of information about these startups improves significantly after 2008. It 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 ar-

ticles covering investments, product launches, and other key company events. Finally, Crunchbase also provides limited information on the founding team, executives, and board members.[10] This data is particularly reliable for companies that have raised funding. Some of the more successful startups from our sample include DoorDash, Slack, Zoomdata, Medium, Bustle, Duolingo, Paytm, GoFundMe, Zomato, Patreon, Coursera, ZipRecruiter, ResearchGate, Giphy, BetterTaxi, RoyalFurnish India, edX, GetHuman, Auth0, and Thrive Market. These companies' business models and the markets they target overlap with the larger sample of firms in our data. In Table A2, we present the broad product categories represented in our sample. As we expected, these are representative of a full range of software-driven technology startups.

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 more than 30,000 web technologies for 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 the Crunchbase startup data. BuiltWith tracked when each technology was first detected 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 these companies. Using this data, we can identify when a company adopts A/B testing technology into its stack.

[Figure 1 about here.]

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

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

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 benchmark performance. We use the SimilarWeb API to pull down weekly website performance metrics for the Crunchbase startups in our sample from the first week of March 2015 through the last week of March 2019.

Internet Archive's Wayback Machine is a non-profit archive of websites on the internet. The Wayback Machine archives hundreds-of-billions of webpages per year, saving the front-end code, website copy, and images. Given the selective archiving of more prominent websites by the Wayback Machine, we focused on pulling historical website snapshots for the 13,012 startups that had raised funding at the start of our sample. These startups are the most likely to be linked to, have prominent profiles, and be regularly archived. Using the Wayback Machine's API, we then pulled a snapshot—when available—at the monthly level for each startup starting in April 2015 and ending in November of 2018.[11]. We were able to pull multiple-snapshots from 8,268 firms with an average of 11 monthly-snapshots per firm, roughly a snapshot every four months. We use this data to capture the nature and magnitude of changes to each page's front-end code-base, which allows us to test if the adoption of A/B testing by firms is associated with substantial changes to a startup's website.

Sample construction

We link startups across these data sources through website URLs. Unlike firm names, URLs are unique identifiers, eliminating the need for fuzzy matches.[12] To ensure that our sample begins with "active" startups at risk of adopting A/B testing tools, we

[11]We started the API pull in December of 2019 to acquire both the code and screenshots of archived websites. However, we ran into several technical difficulties in trying to pull both types of data. In particular, the Wayback Machine's responses sometimes returned complete code and sometimes failed to render the page. We successfully pulled code snapshots after debugging, but not screenshots, because screenshots relied on having the content of image files that were very sparsely archived by the Wayback Machine. The final API pull was run in February of 2019.

[12]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 startup develops and grows its product(s), not corporate structures. For these reasons, we treat each startup URL as an independent firm.

include only startups in the Crunchbase data with 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.[13] 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 associated with my-awesome-blog.wordpress.com). To address this problem, we exclude pages with over 80 active unique technologies as of March 2015.

After these exclusions, our dataset consists of 35,262 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 March 24th, 2019, resulting in 208 weeks (4 years) and a total of 7,334,496 million firm-week observations.

Variable construction

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

Independent and Dependent 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.[14]

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

[14] 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 measurement tools. 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 and are labeled as not doing so.

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 over 60% of the weeks in which firms are A/B testing. The next most prominent A/B testing software is Google, with just over 20% of the market. The remaining 20% is split between Visual Website Optimizer, Adobe, AB Tasty, and Experiment.ly. We provide further details on the heterogeneity of adoption of A/B testing in Table 1.

[Table 1 about here.]

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 5 technologies (5th percentile), while others grow in complexity reaching 111 different web technologies (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.

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

Analysis and Results

We organize our analyses into three parts.

First, in Section I, we use our sample of 35,262 startups over four years to estimate A/B testing's effect using standard two-way fixed effect (TWFE) and event-study type models. Our large number of firms and long panel allow us to estimate a variety of models to check robustness, explore how A/B testing's effect might vary over time, and investigate if there is heterogeneity in where A/B testing appears to matter most.

In Section II, we explore if A/B testing drives incremental or more radical changes. To do so, we use variation in the likelihood that a startup "scales" or "fails," whether the startup launches new products, and how a website's codebase shifts after firms adopt A/B testing tools.

In Section III, we use the March 2017 launch of a new A/B testing tool, Google Optimize 360, as a shock that allows us to identify further the causal effect of A/B testing on startup growth. We use this shock in two ways, both across and within firms. First, to construct an instrument for Google 360 adoption, we use the fact that firms using Google's Tag Manager (GTM) software adopt Google 360 at much higher rates. This is because Google 360 strongly recommends—though it doesn't require—firms use Google's Tag Manager (GTM) software to get the most out of Google 360. Thus, startups that already use GTM should be more likely to adopt Google 360 since 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 360, adopt at twice the rate than non-GTM users. This stylized fact, combined with fixed effects, creates an effective difference-in-differences instrumental variable identification strategy. In essence, we use whether a startup had adopted GTM a year before the launch as an instrument that increases the likelihood a firm utilizes Google 360 while controlling for firm and time fixed effects.

Next, we use the Google 360 shock by leveraging our extended panel to build synthetic controls for each of the Google 360 adopters. We have over 100 weeks of data before the launch of Google 360, during which no startups can adopt since the tool was not available. These data allow us to match growth trajectories using twenty-four months of pre-treatment trends before any firm adopts Google 360. Using these synthetic controls, we then estimate what a startup's growth trajectory would have been had it not started using the Google 360 A/B testing tool.

I. Does A/B Testing impact startup performance?

We begin by assessing the impact of A/B testing on growth by estimating a two-way-fixed-effects (TWFE) model:

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

Our interest is in β, the impact of A/B testing adoption on startup performance (Y_{it}). To reduce selection bias, the model includes 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, and 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 startup idea, the presence of a strategy, location, and founders' educational background, among other fixed resources or capabilities of the startups.

In addition to our fixed effects, we include a weekly time-varying control variable for the number of other technologies adopted by the startup. 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).

Model 1 in Table 2 estimates the raw correlation between the use of A/B testing and weekly visits after accounting for week fixed effects. The relationship is significant, and we see that firms that use A/B testing have 296% more visits than those that do not. In model 2, we include technology stack controls. If the adoption of A/B testing is correlated with the adoption and use of other technologies (e.g., payment processing tools), then 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 significant. 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 significant, with a magnitude of 13%. The point estimate suggests A/B testing improves startup performance.[15]

[Table 2 about here.]

Alternative specifications of the impact of A/B testing

To further assess the robustness of our results, in Figure 2, we present estimates of the effect of A/B testing using a variety of different modeling choices.[16] The left-most estimate (the circle) presents the estimate from model 4 Table 2. 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 where 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 will never adopt A/B testing tools, but all post-failure observations will still be included. By excluding zeros, we end up with an unbalanced panel where observations are treated as censored if they fail. The estimate remains unchanged. It does not appear that the over-representation of zero observations drive our findings.

[Figure 2 about here.]

The next three coefficients repeat the same pattern and include an additional firm-week slope fixed effect for each startup. Including slope fixed-effects allows us to address

[15] In Appendix A1 presents placebo estimates to check if our preferred modeling strategy (column 4) mechanically leads to positive effects. To rule out this possibility, we test whether adopting cloud font tools impacts growth. While we expect larger and faster-growing firms to utilize cloud font tools, we do not believe they should increase startup growth in any meaningful way. We find that once we include our technology stack control and firm fixed effects, the estimated impact of cloud font tools is precisely estimated zero. This suggests our results are not merely a mechanical result of our modeling strategy.

[16] In Appendix Section A2 we extend the robustness checks in Figure 2 by showing that changes to our measure of A/B testing do not alter our findings. In particular, we test if we see the same results when using a more expansive definition of A/B testing software that includes tools that allow for A/B testing, but whose focus is on more general web analytics. We find a very similar pattern of results.

the possibility that our finding is a consequence of fast-growing startups adopting A/B testing at higher rates. While the estimate shrinks to between 5% and 10%, it remains significant. The next set of three coefficients also control for variability in growth rates but instead includes the lagged value of the dependent variable. The effect of A/B testing remains statistically significant and is similar to the slope fixed-effects estimates.

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 maybe its effect is gradual and compounds with time from adoption. Our TWFE results represent the average difference (within firms) between observations where A/B testing is used and firms-weeks where A/B testing is not. Given our panel runs for four years (208 weeks), this implies that the TWFE estimator is an averaging over firms that just adopted, firms that have been using A/B testing for years, and firms that have stopped using A/B testing for equally as long. In essence, the TWFE can be thought of as an average of a multitude of "2-period" (e.g., week one vs. week 26, week one vs. week 52, ...) by 2-group (i.e., using A/B testing vs. not using A/B testing) comparisons (Goodman-Bacon, 2018).

To test if the impact of A/B testing increases over time, in the column "TWFE 2x2" in Figure 2, we show three estimates from models that include two periods of our data. The first estimate (the triangle) comes from a simple difference-in-differences model using only data from the 1st and 26th week or 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% to 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 week 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 impact of A/B testing over time

The last 12 estimates in Figure 2 replicate the first 12, but only include startups that transitioned to or from A/B testing.[17] These "event study" models rely exclusively on comparisons between firms that will or have adopted, increasingly the comparability between firms in our sample. The estimates are somewhat smaller, but still significantly higher 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.

[Figure 3 about here.]

The top panel in Figure 3 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 is little in the way of pre-trends, and the estimates overlap with zero. There are many reasons why we might expect a lack of observable pre-trends. If adoption is driven by sales cycles, this might lead firms to adopt A/B testing tools at points in time unrelated to other events in the firm. If an A/B testing provider significantly raises the tool's price, which firms like Optimizely have done in the past, firms may choose to stop using the A/B testing tool for reasons unrelated to the firm-specific shocks. While we can never rule out unobserved differences in pre-trends with non-experimental data, the lack of any trends before the event-date suggests adoption and dis-adoption decisions may be plausibly exogenous.

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

The effect of A/B testing appears to grow with time, though estimates are noisy. After one month, the impact on growth is around 5%, rising to above 10% after 18 months. Consistent with our "2x2" models above, it appears the value of A/B testing increases overtime.

The bottom panel in Figure 3 shows estimates for a model that includes dummies for 36 months before and after the use of A/B testing. This extended event study again reveals little in the way of pre-trends. While 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%, though the 95% confidence intervals are wide.

II. Mechanisms

How does A/B testing impact startup performance?

We focus our analysis on how A/B testing impacts strategic decision making and product development. While the results so far lend credence to the idea that A/B testing improves startup growth, these findings do not shed light on the changes firms make to achieve this growth. The qualitative insights presented earlier suggest that A/B testing makes launching new products less risky and liberates startups to attempt more changes, introduce more product features, and learn more quickly. Here we test if A/B testing impacts each of these processes.

Scaling and failing

First, to measure how A/B testing impacts firm learning, we test if A/B testing leads to both increased scaling *and* increased failing. A key feature of learning models, especially in entrepreneurship, is that improved learning should lead a firm to discover if it is destined to fail or, if the idea is promising, to more quickly scale (Camuffo et al., 2019; Yu, 2020). To test this prediction, we split our website visits measure into five discrete and mutually exclusive buckets: 0 weekly visits, 1-499 weekly visits, 500-4,999 weekly visits, 5,000-49,999 weekly visits, and 50,000 and more weekly visits.

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, since learning helps startups iterate in their search for product-market fit, we should see firms 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. Again, since these measures are at the startup-week level, we use our primary estimation technique.

Figure 4 shows the estimated effect of A/B testing on whether the firm is more likely to scale or fail. Here, we see a bimodal response to the adoption of A/B testing just as predicted. Firms that adopt AB testing are significantly more likely to experience a substantial increase in visits and zero visits. This suggests that A/B testing firms may be learning faster, both in terms of 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 startup accelerators improve learning, leading to increased failing and scaling (e.g., Yu, 2020).

[Figure 4 about here.]

Incremental and radical product changes

Second, we explore if A/B testing leads the firms to make more website and product changes. We do this in two ways. As mentioned above, we use the Internet Archive's Wayback machine to extract monthly snapshots of the code that generates each startup's homepage. We focus only on startups 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 startups with multiple observations. Using this data, we test if firms using A/B testing in a month t differ 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 $\Delta_i t$ 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 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 the fact that larger websites and a longer duration between snapshots will exhibit more website change. That said, our pattern of results does not shift if we exclude these two controls.

We then calculate four different measures for how the website changes. We count the (logged) total number of code lines that changed between t and t'. While crude, changes to a code-base have proven to be a useful proxy for how much a website or digital product has shifted (e.g., MacCormack, Rusnak and Baldwin, 2006). Second, we use an algorithm developed by Gowda and Mattmann (2016) that compares HTML tree structures returning a dissimilarity score that that ranges from 0 (the same HTML tree) to 1 (completely different). Third, we use a related algorithm that compares the similarity in the website's style (technically the website's CSS code) to test if A/B testing firms are merely shifting the shade of blue or make more meaningful design changes. Finally, to get a sense of whether A/B testing firms are making incremental changes to their code or more radical changes to their online presence, we also dichotomize our measure of lines changed into changes in the top 5% of the distribution (e.g., significant changes.)

While the Wayback machine allows us to test if A/B testing changes how firms shift their code, it is very much an imperfect measure of a startup's product development process. To gain further insight into how A/B testing impacts product development, we measure the number of products a startup has launched during our panel. 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 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 Appendix Section A4 for additional examples of articles that did and did not get tagged as covering product launches. Since 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 startup is engaging in new-product idea generation. Since all these measures are at the startup-week level, we use our basic model to test if A/B testing improves these metrics.

Figure 5 shows how A/B testing impacts the startup's product development process.[18] The first estimate shows the estimate on the number of lines changed between each Wayback machine snapshot. The estimate suggests that 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 non-experimenting firms. The 4th row shows whether A/B testing firms are just making more, but smaller, changes to their website by looking at the probability of a significant code change. If this were the case, we would expect the average number of code lines changed to be higher, but that A/B testing firms would be less likely to make significant changes. However, this is not the case. The final estimate suggests that A/B testing firms launch more new products as measured by CrunchBase news articles. On average, a startup 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

[18]In our Appendix Section A3, we report the regression models that correspond to Figure 5. We also report models testing if A/B testing improves additional measures of product-development and fit. Again we find positive effects.

launched 0.36 products. Our estimate implies that an A/B testing firm has a risk of launching a product in a given week that is 18.6% greater than the average firm.

[Figure 5 about here.]

Heterogeneous effects of A/B testing

How does the impact of A/B testing vary across different startup types? In Figure 6, we present the produces estimates across different sub-samples of our data using our baseline specification from Table 2 model 4. Instead of including interactions with startup characteristics, which can be challenging to interpret in fixed-effects models (Shaver, 2019), we instead estimate separate models for the indicated sub-sample.

In Panel A of Figure 6, we analyze the heterogeneity in our effect based on startup characteristics—including factors such as venture funding, age, location, as well as size. Statistically, these estimates are broadly similar across the specifications, with most estimates being positive and statistically significant and ranging from a 5% effect to a 20% effect. The standard errors across the various subsamples overlap, and statistically, most of this heterogeneity is not significant. However, in terms of point estimates, we see younger startups outside the Bay Area that are smaller and lack venture capital funding having somewhat more significant A/B testing effects. More early-stage startups that are not part of established entrepreneurial networks may learn faster using A/B testing.

In Panel B, we analyzed heterogeneity in the impact of any testing based on a startup's primary industry category. Here too, we find little heterogeneity in effect sizes across most industry categories. Effect sizes again range from 5% to 20% but appear most significant for startups focusing on e-commerce, software, financial services, education, healthcare, and hardware. These estimates suggest that A/B testing is likely to improve growth across a wide range of different startups.

[Figure 6 about here.]

III. Additional robustness tests

The launch of Google Optimize 360

While the evidence thus far points to A/B testing improving startup 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. While our models account for many forms of selection bias, the specter of unobserved time-varying shocks remain. Second, our TWFE and event study estimates pool together many different tools and decisions to adopt and dis-adopt. While doing this increases our statistical power and our results' generality, it also means that our estimates mix potentially varying effects. Thus, while the panel data analysis above provides evidence for the impact of A/B testing at a high-level, it sidesteps many details that are valid causes of concern.

To address these concerns, we focus on the global launch of Google's Optimize 360 A/B testing suite (Google 360) on March 30th, 2017. The tool, which integrates on top of Google Analytics, is similar to competitors like Optimizley 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 talent support. The Google 360 suite offers both a free and paid tier. 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 startup performance.

Instrumental variables

While the choice to adopt Google 360 after it launches is endogenous, we construct an instrument for adoption by noting that Google 360 strongly suggests firms also use Google's Tag Manager (GTM) technology. GTM allows startups to manage website "tags" without writing any code. These tags are used for search engine optimization and analytic pipelines.

We focus on startups that adopted GTM at least a year before the launch of Google 360. These startups almost certainly adopted GTM without knowledge of whether, and

especially when, Google would launch a new A/B testing tool. This suggests that, after accounting for firm fixed effects, we can instrument Google 360 adoption by whether the firm had adopted GTM at least a year earlier. Formally, we fit a difference-in-differences instrumental variable model where the first stage is:

$$Google360_{it} = \pi Z_{it} + \phi(Technology\ Stack_{it}) + \omega_i + \psi_t + \eta_{it} \tag{3}$$

where the instrument $Z_{it} = GTM_{it} \times PostLaunch_t$ and the second-stage is:

$$Y_{it} = \beta(Google360_{it}) + \theta(Technology\ Stack_{it}) + \alpha_i + \gamma_t + \epsilon_{it} \tag{4}$$

In both equations, we include startup and week fixed effects and our time-varying technology stack control. This model relies on two critical assumptions. First, after controlling for firm fixed effects and technology stack controls, GTM status is exogenous. This seems plausible given that firms that adopted GTM at least a year before the shock are unlikely to have insider knowledge of Google's future moves. Second, we must assume an exclusion restriction that GTM doesn't shift post-shock startup performance in other ways. Since we include time and firm fixed effects in our models, any violation implies that GTM's performance effects suddenly changed for other reasons right when the Google 360 tool was launched. The latter seems unlikely.[19]

We estimate our difference-in-differences instrumental variable model on the sample of startups at risk of adopting Google 360 at the time of the launch. We do so by only including startups with the Google Analytics tool—a requirement for using Google 360 and GTM—installed in the week before launch. Over 90% of startups in our data use Google Analytics, and those startups that do not are much more likely to be failed ventures with websites with few to no technologies. We also only include startups with more than 0 page views at the Google 360 launch in March 2017. By

[19]Crucially, it does not appear that Google launched any significant new GTM features or integrations in the year before or after the Google 360 product launch. Reading through the product release notes https://support.google.com/tagmanager/answer/4620708?hl=en reveals in 2017 the major product changes (other than improved integration with Google 360) are GDPR compliance measures, bug fixes, changes to the UI, and a handful of new tags being launched.

excluding these startups, we eliminate most startups that had failed before the launch and so mechanically cannot adopt. Our final sample restriction is that we focus only on startups that had yet to choose any A/B testing tools before the launch. Such a limit ensures that our estimates are not picking up switching between A/B testing tools, but instead the impact of adoption. This restriction brings the total number of observations from firms in our sample from 35,262 to 20,958.[20]

We begin our instrumental variables analysis by checking that GTM use predicts whether a startup adopts Google 360. In Figure 7, we plot the percent of startups who have installed Google 360 against the number of weeks relative to launch. The dashed line is the adoption rate for startups that had installed GTM at least a year before the launch (12.76% of our sample); the solid gray line depicts firms that did not have GTM installed. While adoption rates are low, 2% to 5% by the end of our panel, startups using GTM appear two to three times more likely to adopt Google 360, suggesting a strong first stage.

[Figure 7 about here.]

Table 3 presents our IV estimates and the first stage F-statistics (Appendix A4 reports the first-stage regression results). Model 1 shows the second-stage estimate (equation 4 above). The estimate is positive and significant. However, the estimate, 5.88, is an order-of-magnitude greater than even our largest TWFE estimate.

There are likely several reasons for this difference. First, our instrumental variable approach estimates the effect for firms that complied with the instrument (i.e., it estimates the LATE). Our TWFE estimates the impact for any firms that start or stop A/B testing. Indeed, Figure 6 shows that the effect of A/B testing is more significant for smaller, younger, and less established firms. If compliers are more likely to be these less established firms, then it seems plausible that adopting Google 360 could help a 1,000 visits a week startup scale to have anywhere from 2,600 to 9,159 visits a week,

[20]There was an invite-only beta program for Google Optimize 360. In our sample, 23 firms are recorded as having Google 360 installed two months before the launch date, presumably because they were part of the beta testing program. We drop these firms from our analysis. That said, including these 23 firms does not impact our results.

the range of our 95% confidence intervals. While such magnitudes seem unlikely for a firm that already has hundreds of thousands of visitors, if the LATE excludes them, it might explain our estimates.

[Table 3 about here.]

The second reason our IV estimate is substantial is that, while our first stage is strong (the F-statistic in our models are above 40), the absolute magnitude of the difference in adoption rates is relatively small. Indeed, while the first stage model suggests GTM startups are 72% more likely to adopt, a sizeable relative effect represents a shift in adoption rates from 2.49 to 4.29, a 1.8 percentage point effect. This fact introduces instability in instrumental variable models because the second stage estimate, our β, relies on dividing the reduced-form estimate by the first stage effect π. This fact is seen in the canonical Wald-estimate for β in a just-identified instrumental variable model with outcome Y, binary treatment D, and a binary instrument Z:

$$\beta = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} \tag{5}$$

Notice that, in this simple case, the denominator is merely the difference between uptake in the instrumented and non-instrumented groups. In our case, the denominator is similar to this version but takes into account fixed effects and controls. Our first-stage estimate suggests the difference in adoption between the two groups is about two percentage points, suggesting that our IV estimate will scale the reduced-form numerator roughly 50 times (1/0.02). Given the low adoption rates of Google 360, there is only so much we can do to address this problem. Because our denominator scales our numerator, our 95% confidence intervals for the impact of Google 360 are broad, ranging from increases of 260% to 915%.

In model 2, we attempt to increase the denominator's size by interacting with our instrument the number of weeks and number of weeks squared since launch. Figure 7 shows that GTM startups saw a spike in early adoptions compared to non-GTM using startups. These interactions could potentially provide us with a more substantial first-

stage gap. Indeed, the point estimate drops to 395%, though it is not statistically different than the estimate in column 1.

Further, the difference-in-difference IV estimator that underlies the results in columns 1 and 2 relies on two strong assumptions (De Chaisemartin and d'Haultfoeuille, 2018). First, it only identifies a LATE if the treatment effects are assumed to not vary over time. Second, it relies on the assumption that the treatment effect is equal in both groups. We find it unlikely that our treatment, the adoption of Google 360, satisfies these two assumptions. First, our event-study and 2×2 results suggest that the magnitude of the A/B testing effect varies with time. Second, it seems plausible that Google 360 is more effective when used with GTM. After all, Google recommends pairing these tools.

Fuzzy difference-in-difference

To address these concerns, we also estimate our effects using the fuzzy difference-in-differences time-corrected Wald ratio proposed by De Chaisemartin and d'Haultfoeuille (2018). This estimate does not rely on stable-in-time treatment effects nor similar treatment effects across groups. However, it does assume that the share of adopters is stable over time in the "control" group (e.g., non-GTM using startups). As Figure 7 shows, this is not the case. Fortunately, De Chaisemartin and d'Haultfoeuille (2018) develop a method that partially identifies the treatment effect when the control group is unstable, allowing us to place a lower bound on the estimate. While this fuzzy difference-in-differences method can be extended to multi-period panels, doing so requires additional assumptions. Instead, we estimate these models using only two periods of our data: the week before the launch of Google 360 and the final week in our panel. Beyond simplifying estimation, doing so has the benefit of making the estimated effect easier to interpret. It is the effect on weekly visits after two years for firms pushed to adopt Google 360 by our instrument during this period.

Column 3 in Table 3 shows the point estimate from the fuzzy-DiD model, making the unrealistic assumption that the non-GTM control group is stable in the adoption

rate. The model includes startup and time fixed effects along with our technology stack control. As with models 1 and 2, the first stage F-statistic is strong. The estimated effect is 321%, somewhat smaller than the estimates in columns 1 and 2, though not statistically different from our prior estimate. In column 4, we drop the stable treatment group assumption, but at the cost that the estimate is now the lower bound for a potential effect.

Further, the bounded estimator does not allow us to include any controls. Regardless, the lower bound is similar in size to the first three columns at 346%. Overall, the fuzzy-DiD estimates suggest that our estimates are not an artifact of assuming constant treatment effects across time or treatment groups.

TWFE using Google 360

To further check the validity of our instrumental variable findings, we also estimate the effect of Google 360 using standard TWFE models like in model 1 of Table 2 but with the sample used to estimate our instrumental variables. The estimate in model 1 in Table 4 is 33.8%, still larger but closer in magnitude to our TWFE estimates in our earlier analysis. This pattern is consistent with the idea that A/B testing's effect is more significant when only estimated from our sample's compliers. In model 2, we drop non-GTM adopters from the sample to see if GTM using startups, startups that are twice as likely to have been exogenously pushed into adopting, still see a boost. Indeed, the effect remains nearly unchanged. Finally, in model 3, we only keep GTM adopters who adopted within six months of the launch under the theory that these firms were the most likely to be "pushed" into adoption. Again, the estimate remains positive, significant, and near 30%.

[Table 4 about here.]

Finally, in model 4, we again utilize a 2 $\times$ 2 type-design to test the robustness of results and check for trends in our treatment effect. In this model, we only include the week before the launch and the week at the end of our panel, two years after the

launch of Google 360. After two years, we see a substantially larger estimate at 0.789. One possibility is that variation in the treatment effect size can explain why the IV estimates are much larger than the TWFE estimates. In the next section, we more formally test if A/B testing's impact increases with time.

Synthetic controls

Our final robustness check uses the fact that we have two years of weekly data before the launch and adoption of Google 360 to fit synthetic control models. Using the sample from our instrumental variables models, we use the 105 weeks of pre-launch visit data to build synthetic controls for each of the Google 360 adopters. Crucially, since no startup can adopt Google 360 during this period, we can be confident that early Google 360 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 startups not adopted the Google 360 A/B testing tool. An additional benefit of this approach is that, since we construct time-varying counterfactuals, we can improve on our event-study and 2x2 estimates to directly test if the effect 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, Diamond and Hainmueller (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 non-parametric 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 8 shows the estimated treatment effect relative to the time of Google 360 adoption. Panel A focuses on the effect trajectory a year before and after adoption; Panel B shows the full set of estimates which run two years before and after. The fact that the estimates are near zero before adoption suggests that the model generates

precise pre-adoption counterfactuals.[21]

[Figure 8 about here.]

Turning to the post-adoption 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 A/B testing's impact increases with time since adoption, though Panel B indicates that it stabilizes a little after one year.

Table 5 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 pre-trends. By 26 weeks (column 2), the estimate is 37% and significant at the 5% level, though 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 post-adoption period. The estimate is 67.6% with 95% confidence intervals ranging from 24.6% to 93%.

[Table 5 about here.]

Beyond providing further robustness checks, the synthetic control analysis results help explain the variation in effect sizes we see with the 2x2, 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 will lead to more long-term testers who 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 include 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

[21] In Appendix Section A5 we show the growth trajectory of 6 startups and the estimated counterfactual for that startup. The model appears to capture varied growth trajectories at an appropriate level of detail.

relatively seamlessly with the GTM tool they already rely on, we should expect the IV estimates to be more significant. 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 A/B testing improves startup performance and increasingly so with time. Moreover, this technology and approach appear to enhance the performance of a wide array of startups, including those inside and out of entrepreneurial hubs and those in a variety of industries ranging from e-commerce to education.

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 weekly performance measures.

We find that A/B testing is associated with a 5-15% increase in website visits and is positively correlated with increased product introductions, code changes, and other performance metrics. 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. Overall, we find little evidence of large differences in the benefits of technologies across different kinds of startups. The most pronounced difference is among younger and older startups, with younger startups appearing to benefit more.

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 experimentation and exploration approaches suggested by (Knudsen and Levinthal, 2007; March, 1991). We offer a novel insight highlighting an important distinction between running *an* experiment versus a strategy based on *experimentation*. Running a single test will most likely 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 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 TWFE specifications, we include controls for the adoption of other technologies, a rich suite of fixed effects, and heterogeneous firm-level growth trajectories. Our data's dynamic panel structure,

coupled with these controls, allows us to rule out various kinds of selection bias, reverse causality, and omitted variable bias.

Moreover, we provide significant additional analyses, including IV estimation, synthetic controls, and a fuzzy difference-in-difference estimation, to test our findings' robustness. In all our specifications, our results remain significant and consistent. Our conclusions, demonstrated on multiple performance metrics, are also in line with previous studies on the effect of data-driven decision making on firm performance (Brynjolfsson, Hitt and Kim, 2011).

That said, we do not observe the A/B tests that startups 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 a low cost. The cost of experimentation is still high in many other industries, though innovations like 3-D printing may be changing this dynamic. Further, while we consider the long-term impact of A/B testing, we cannot 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. Finally, our findings cannot discern whether A/B testing is part of a broader research and development program. Future research should investigate this phenomenon more deeply to understand better which organizational structures are most aligned with an experimental strategy.

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 an individual organization's benefits? Finally, how will experimentation across the economy change the types of innovations that firms develop and how they are distributed? Addressing these questions, among others, will guide future research and practice.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American statistical Association* 105(490):493–505.

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.

Boulding, William, Eunkyu Lee and Richard Staelin. 1994. "Mastering the mix: Do advertising, promotion, and sales force activities lead to differentiation?" *Journal of marketing research* 31(2):159–172.

Browne, Will and Mike Swarbrick Jones. 2017. "What works in e-commerce-a meta-analysis of 6700 online experiments." *Qubit Digital Ltd* 21.

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 R Fabrizio. 2014. "Using users: When does external knowledge enhance corporate product innovation?" *Strategic Management Journal* 35(10):1427–1445.

Chatterji, Aaron, Solène Delecourt, Sharique Hasan and Rembrand Koning. 2019. "When does advice impact startup performance?" *Strategic Management Journal* 40(3):331–356.

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

Cohen, Wesley M, Richard R Nelson and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science* 48(1):1–23.

Dahlander, Linus and Henning Piezunka. 2014. "Open to suggestions: How organizations elicit suggestions through proactive and reactive attention." *Research Policy* 43(5):812–827.

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.

De Chaisemartin, Clément and Xavier d'Haultfoeuille. 2018. "Fuzzy differences-in-differences." *The Review of Economic Studies* 85(2):999–1028.

Denrell, Jerker. 2003. "Vicarious learning, undersampling of failure, and the myths of management." *Organization Science* 14(3):227–243.

Denrell, Jerker and James G March. 2001. "Adaptation as information restriction: The hot stove effect." *Organization Science* 12(5):523–538.

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

Dyer, Jeffrey H and Nile W Hatch. 2004. "Using supplier networks to learn faster." *MIT Sloan Management Review* 45(3):57.

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.

Garcia-Macia, Daniel, Chang-Tai Hsieh and Peter J Klenow. 2019. "How destructive is innovation?" *Econometrica* 87(5):1507–1541.

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.

Goodman-Bacon, Andrew. 2018. Difference-in-differences with variation in treatment timing. Technical report National Bureau of Economic Research.

Gowda, Thamme and Chris A Mattmann. 2016. Clustering web pages based on structure and style similarity (application paper). In *2016 IEEE 17th International conference on information reuse and integration (IRI)*. IEEE pp. 175–180.

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.

Jeppesen, Lars Bo and Karim R Lakhani. 2010. "Marginality and problem-solving effectiveness in broadcast search." *Organization science* 21(5):1016–1033.

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

King, Andrew A, Brent Goldfarb and Timothy Simcoe. 2019. "Learning from testimony on quantitative research in management." *Academy of Management Review* .

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.

MacCormack, Alan, John Rusnak and Carliss Y Baldwin. 2006. "Exploring the structure of complex software designs: An empirical study of open source and proprietary code." *Management Science* 52(7):1015–1030.

Madsen, Peter M and Vinit Desai. 2010. "Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry." *Academy of management journal* 53(3):451–476.

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

McDonald, Rory M and Kathleen M Eisenhardt. 2020. "Parallel play: Startups, nascent markets, and effective business-model design." *Administrative Science Quarterly* 65(2):483–523.

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.

Mowery, David C, Joanne E Oxley and Brian S Silverman. 1996. "Strategic alliances and interfirm knowledge transfer." *Strategic management journal* 17(S2):77–91.

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.

Ott, Timothy E and Kathleen M Eisenhardt. 2020. "Decision Weaving: Forming Novel, Complex Strategy in Entrepreneurial Settings." *Strategic Management Journal* .

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.

Shaver, J Myles. 2019. "Interpreting interactions in linear fixed-effect regression models: When fixed-effect estimates are no longer within-effects." *Strategy Science* 4(1):25–40.

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.

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

Thomke, Stefan H. 2020. *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Press.

Timmermans, Stefan and Iddo Tavory. 2012. "Theory construction in qualitative research: From grounded theory to abductive analysis." *Sociological theory* 30(3):167–186.

Urban, Glen L and Eric Von Hippel. 1988. "Lead user analyses for the development of new industrial products." *Management science* 34(5):569–582.

Urban, Glen L and Gerald M Katz. 1983. "Pre-test-market models: Validation and managerial implications." *Journal of Marketing Research* 20(3):221–234.

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.

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

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

Table 1: Our panel covers 35,262 startups for 208 weeks (4 years). Panel A provides summary statistics at the startup-week level. Panel B shows the number of startups of each type in our sample and the percent of each type that use A/B testing tools for at least one week over the course of our panel.

Panel A: Startup-week level

	Mean	Median	SD	Min	Max	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: Startup level

	Number of Startups	Percent 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 US	14,645	16.1%
In US, 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%
Below 1,500 Weekly Visits	17,189	8.1%
Above 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%
Health Care	1,042	19.2%
Hardware	1,030	16.5%
Other	12,671	14.2%

Table 2: TWFE regressions show that A/B testing startups experience greater growth.

	(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 Weeks	208	208	208	208
Week FE	Y	Y	Y	Y
Firm FE			Y	Y
Technology Stack Control		Y		Y

Standard errors in parentheses. Brackets show 95% confidence intervals.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Difference-in-differences Instrumental Variable models estimating the impact of the Google 360 A/B testing tool on startup growth.

	(1)	(2)	(3)	(4)
	Log(Visits + 1)			
Using Google 360 A/B tool?	5.88 ***	3.956**	3.207**	3.462**
	(1.67)	(1.402)	(1.055)	(1.196)
	[2.605,9.159]	[1.207, 6.704]	[1.348, 5.529]	[1.358, 6.053]
Model	IV	IV	Fuzzy DiD	Fuzzy DiD
	2SLS	GMM	Stable	Unstable
First Stage F-statistic	46.89	16.54	44.62	44.62
GTM Instrument	Y	Y	Y	Y
GTM × Weeks Instruments		Y		
Observations	4,359,264	4,359,264	41,916	41,916
Week FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	

Error terms clustered at the firm level.

Standard errors in parentheses. Brackets show 95% confidence intervals.

For Models 2-3 standard errors are estimated using a bootstrap with N=500.

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

Table 4: TWFE models estimating the impact of the Google 360 A/B testing tool on startup growth.

	(1)	(2)	(3)	(4)
	Log(Visits + 1)			
Using Google 360 A/B tool?	0.338***	0.352***	0.291*	0.789***
	(0.047)	(0.077)	(0.128)	(0.074)
	[0.246, 0.430]	[0.202, 0.502]	[0.041, 0.541]	[0.645, 0.933]
Observations	4,359,264	4,264,624	4,242,160	41,916
Adopting Sample	All	GTM	Early GTM	All
2-Year 2x2				Y
Week FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y

Error terms clustered at the firm level.

Standard errors in parentheses. Brackets show 95% confidence intervals.

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

Table 5: Generalized synthetic control estimates show that startups which adopt the Google 360 A/B testing tool see increased startup growth.

	(1)	(2)	(3)	(4)
		Log(Visits + 1)		
	1 week before	26 weeks after	52 weeks after	Average
Using Google 360 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 FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y
Number of Factors	5	5	5	5

Standards errors and confidence intervals calcualted using a nonparametric bootstrap with N=500.

Number of factors selected using a cross-validation procedure.

Standard errors in parentheses. Brackets show 95% confidence intervals.

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

Figure 1: Example of the BuiltWith Data

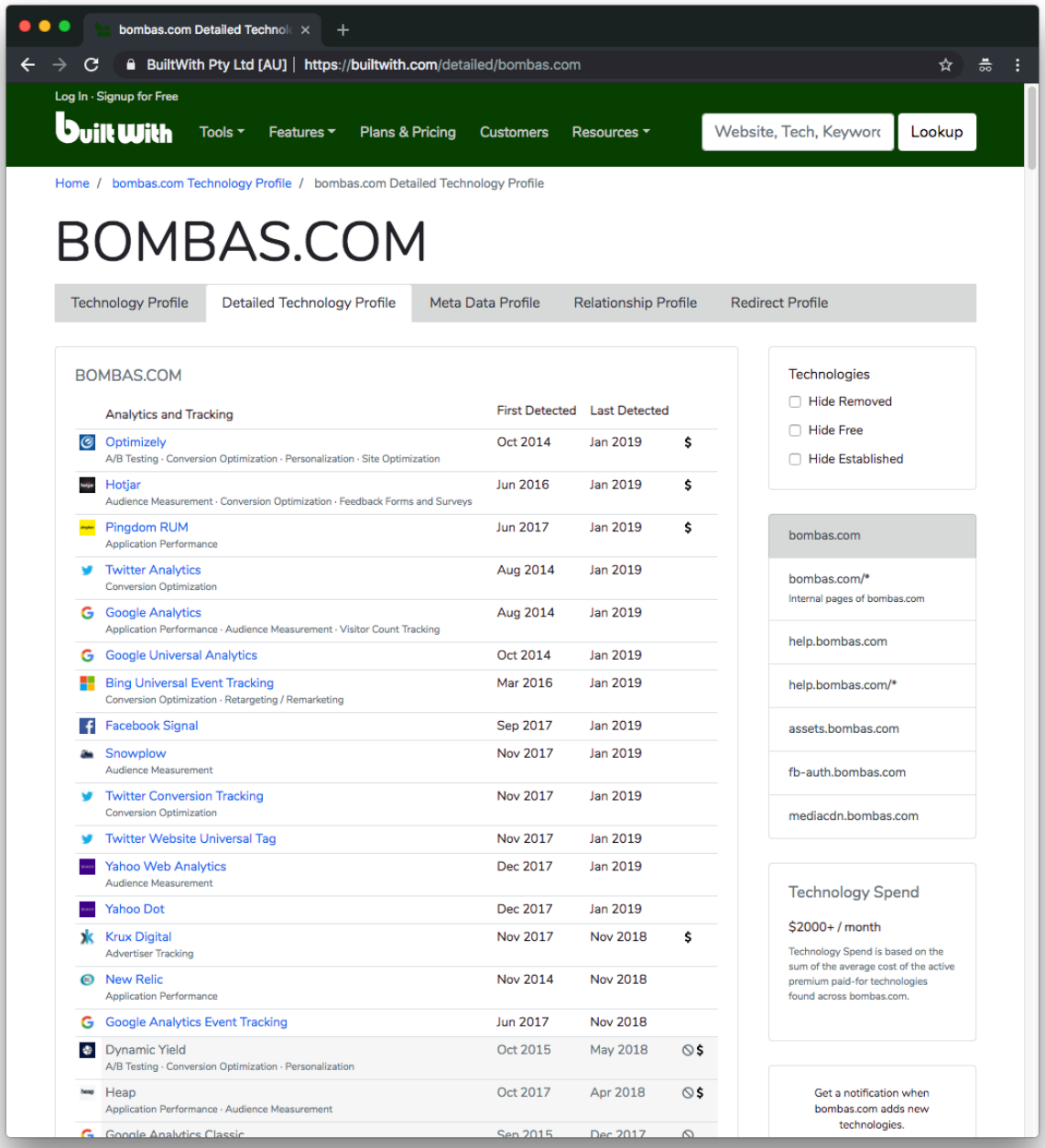


Figure 2: The effect of A/B testing on startup growth holds across a range of model specifications.

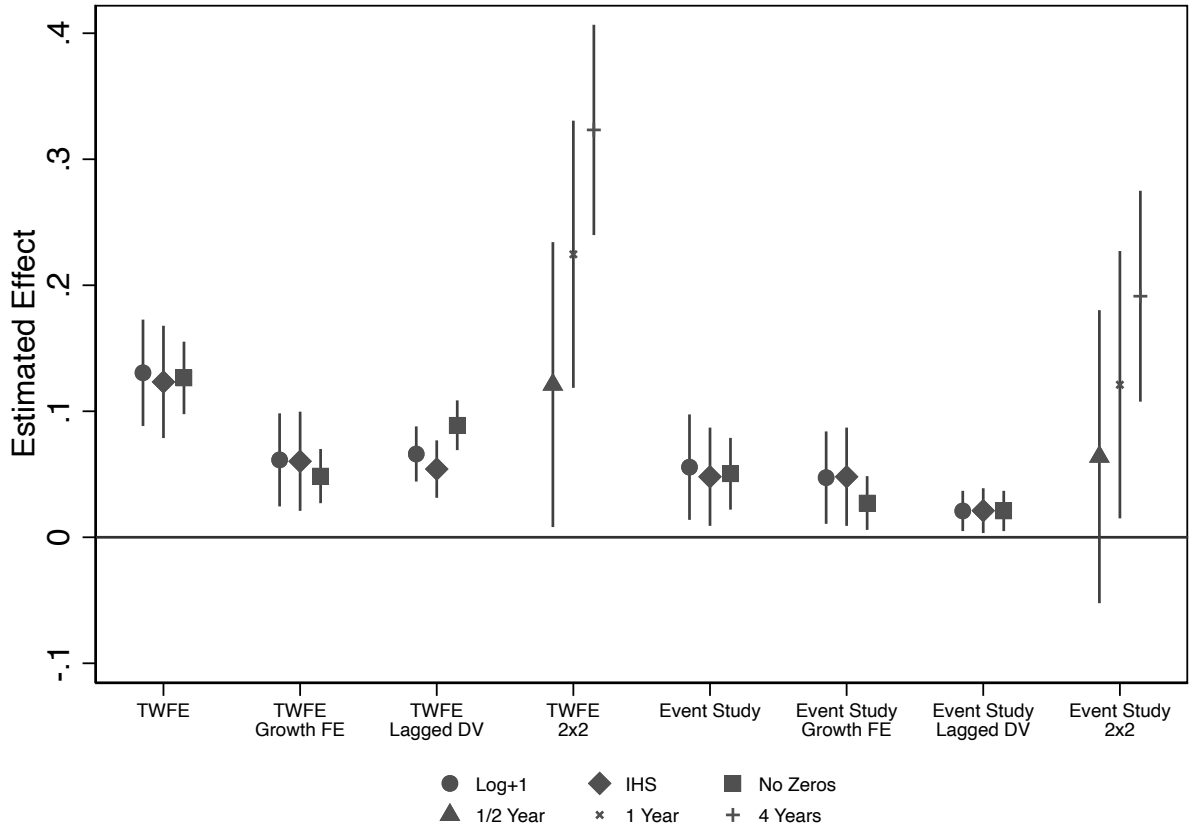
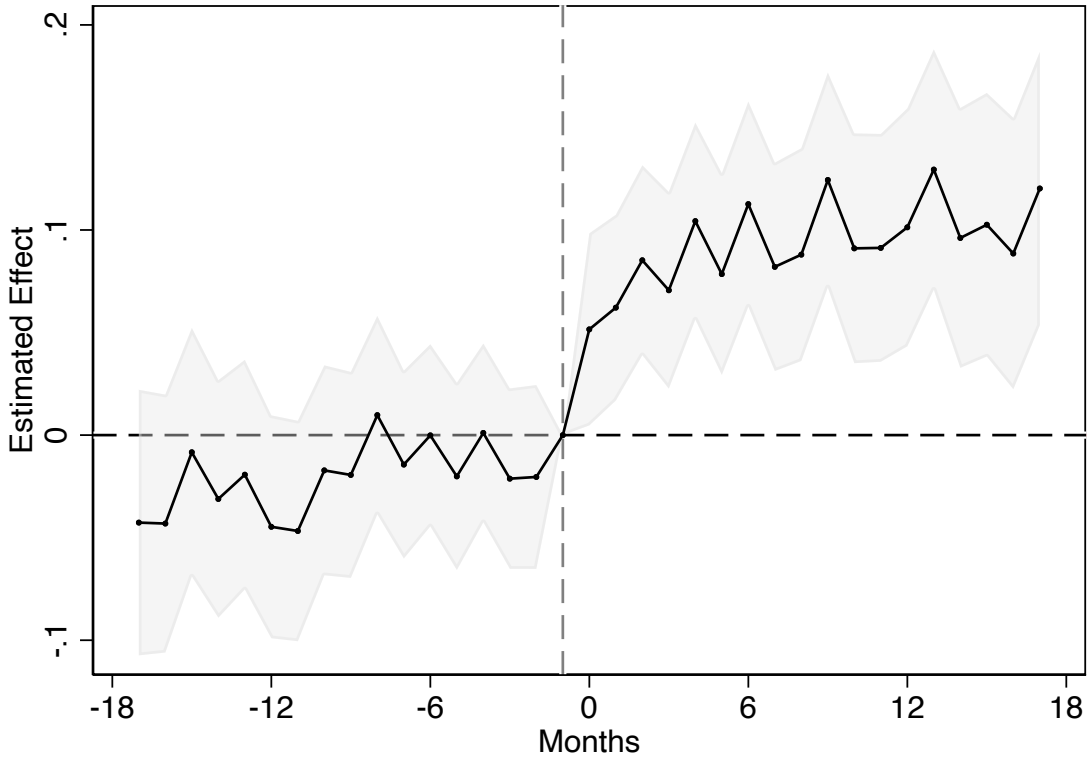


Figure 3: Event Study showing the effect of A/B testing over time. The top panel shows the effect 18 months before and after the event. The bottom panel 36 months before and after the event. The month before the event serves as the excluded baseline.



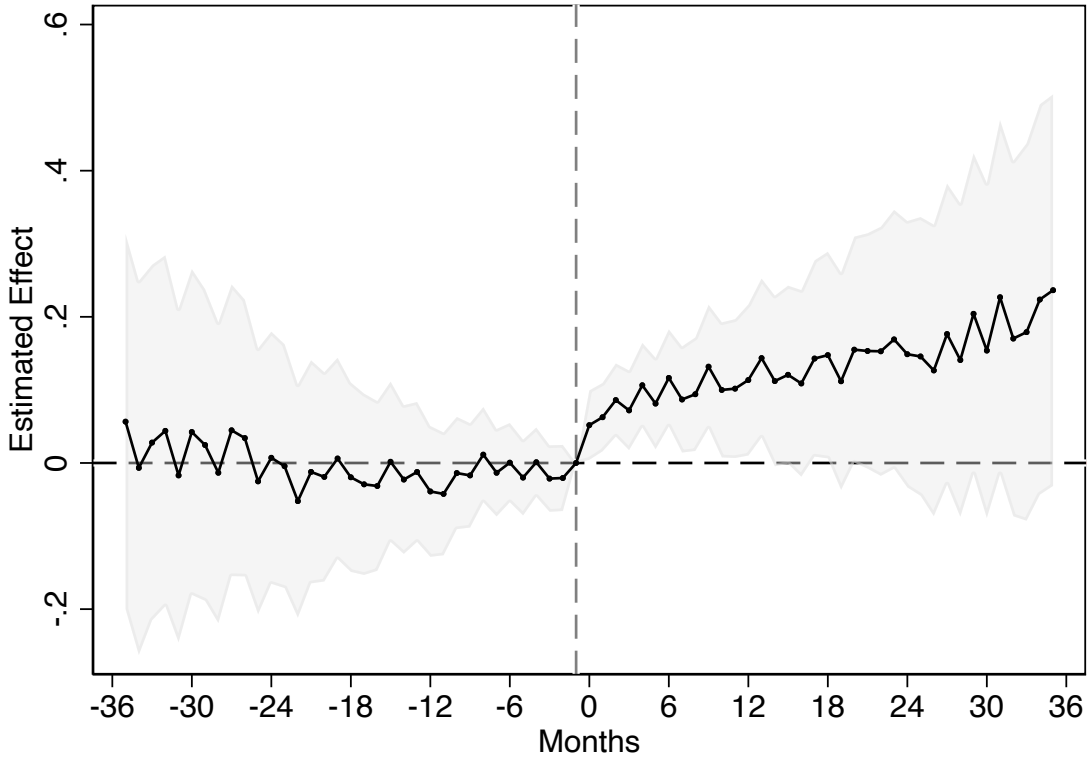


Figure 4: Startups that A/B Test are more likely to fail (0 visits) and scale (achieve more than 5,000 visits). They are less likely to experience middling growth outcomes.

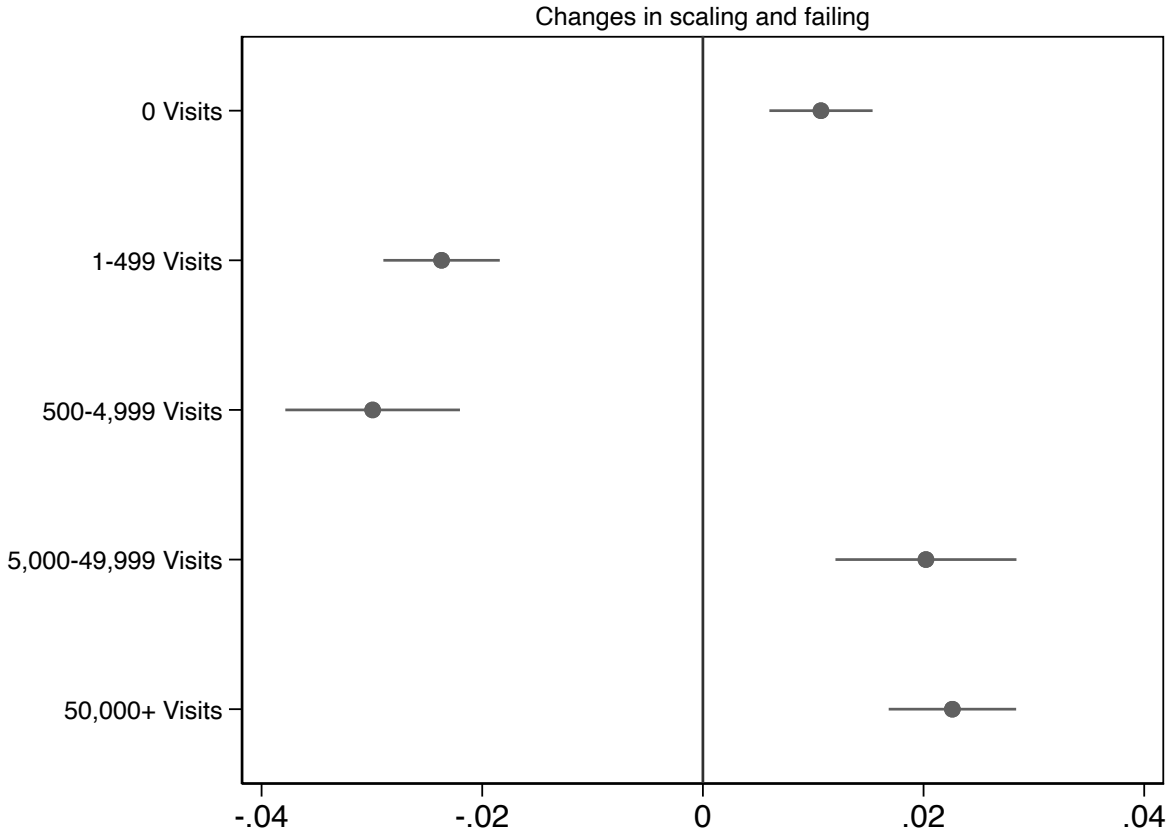


Figure 5: Startups that A/B test make more changes to their website code, make larger changes to the structure of their HTML, make bigger style changes, make large code changes, and launch more new products.

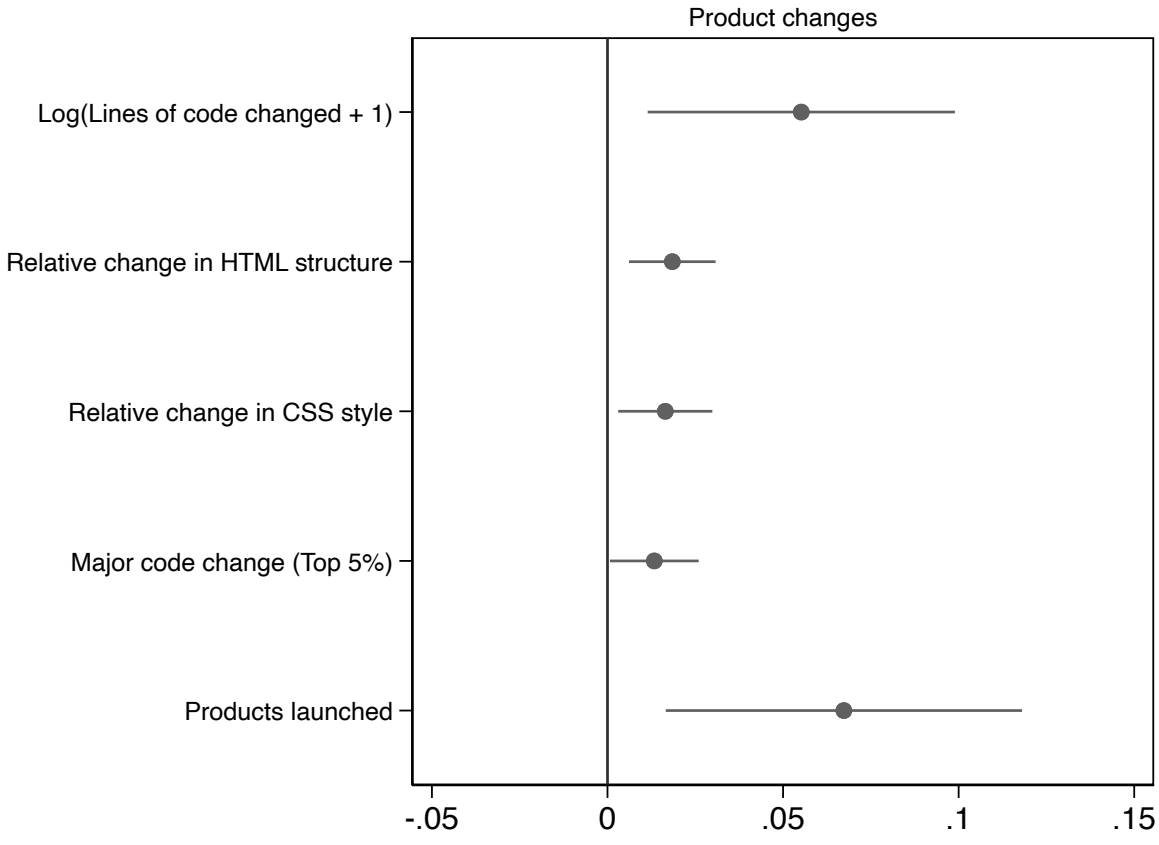
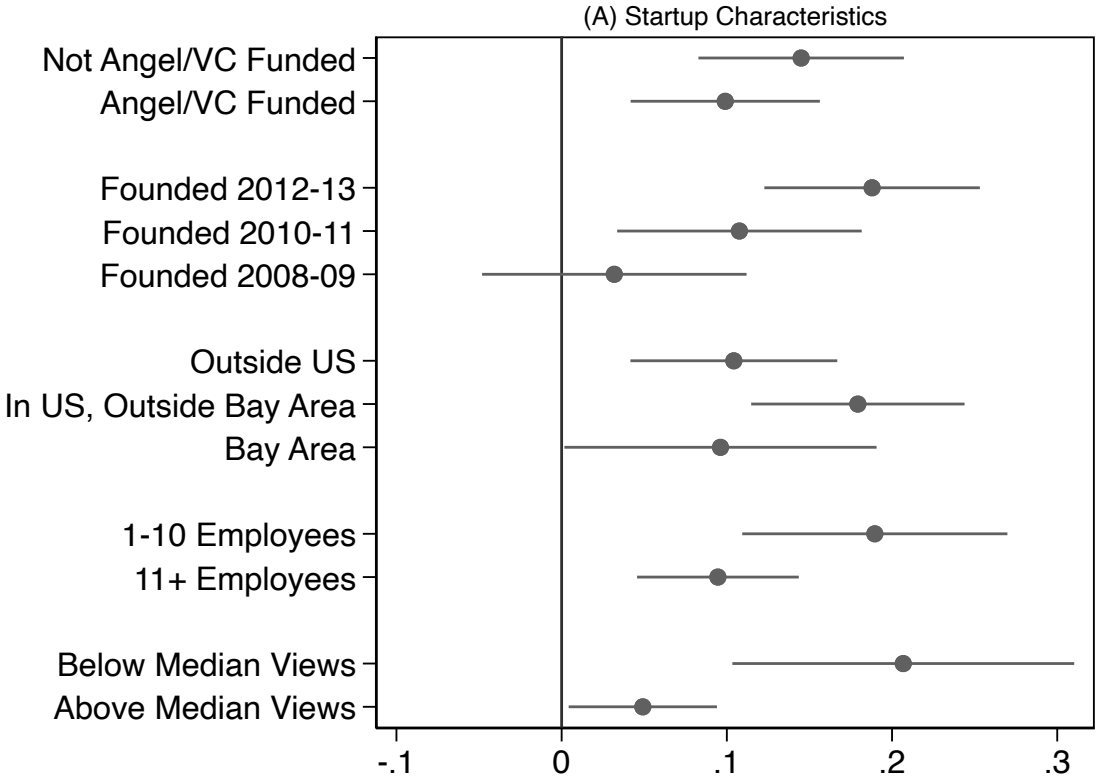


Figure 6: Heterogeneous effects of A/B Testing by startup type (top panel) and industry category (bottom panel). Each estimate represents the effect of a A/B testing on growth for the sample of firms that meet the condition listed. Figure 1 Panel B reports the number of firms of each types.



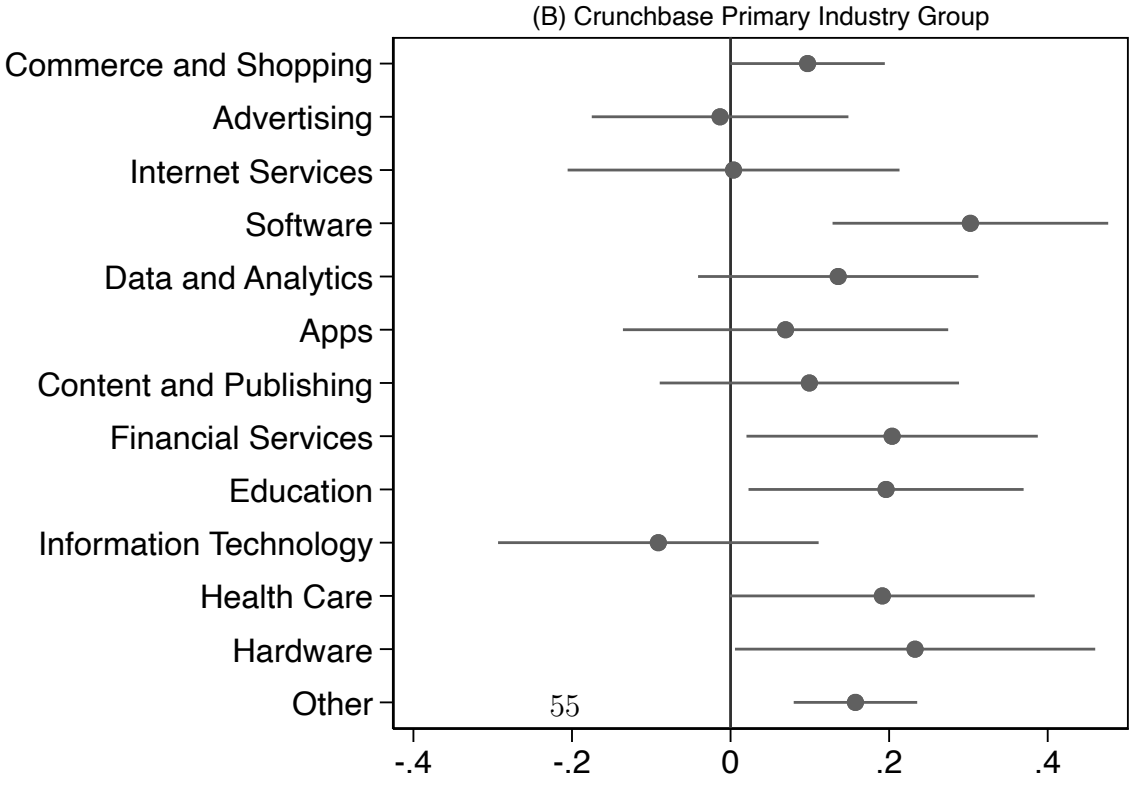


Figure 7: Adoption of Google 360 by whether the startup adopted GTM at least a year before the Google 360 launch. GTM startups adopt at significantly higher rates.

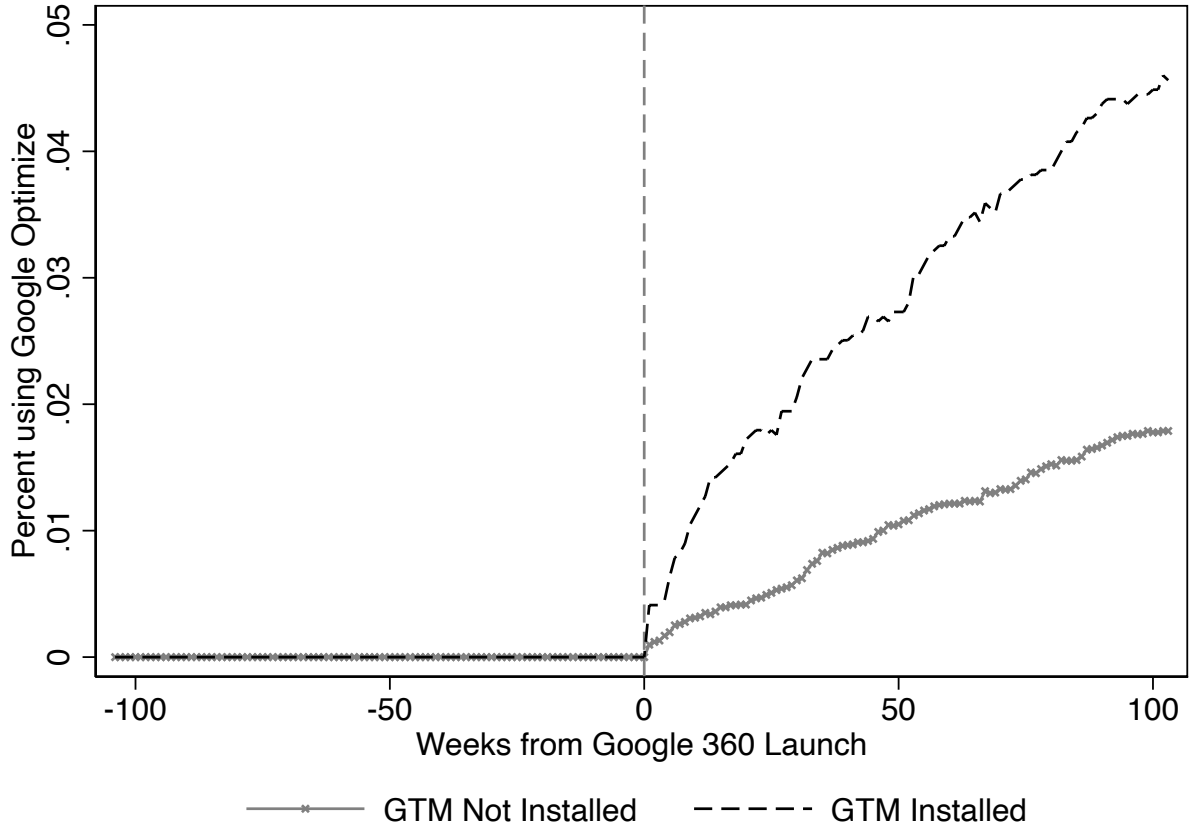
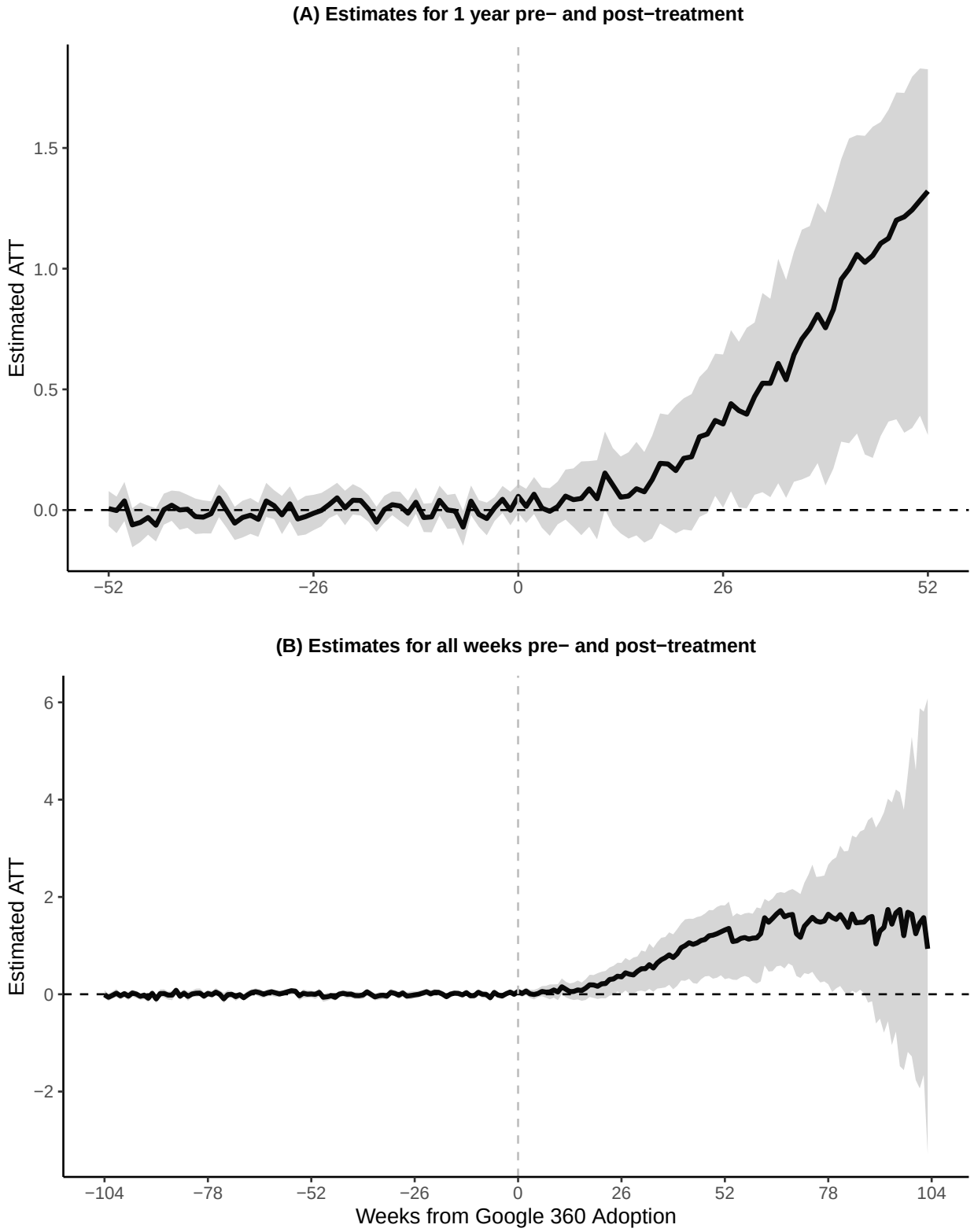


Figure 8: Generalized synthetic control estimates show that startups which adopt the Google 360 A/B testing tool see increased startup growth. The bottom panel 2 years before and after adoption.



A1 Placebo test using font libraries

In Table A1 we run a placebo test by checking if using cloud font providers increases growth. 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.* We then replicate Table 2 in the body of the paper using our font placebo. In Model 1 we find strong evidence for selection. Using a Cloud Font provider is associated with nearly 150% more pageviews! However as soon as we control for our technology stack control the effect turns negative. Our preffered model, with firm fixed effects and our time-varying technology stack control, estimates that cloud font tools have a precisely estimated zero effect. The results in Model 3 highlight the important of not just including firm-fixed effects but also accounting for time-variation in the technologies a startup uses.

[Table A1 about here.]

A2 Impact of A/B enabled tools

In Figure A1, we show our findings hold across a variety of models when using a more expansive definition of A/B testing tools. In the paper's body, we focus on tools that only and explicitly focus on A/B testing. We include both these core A/B testing tools and analytics tools that enable or integrate with A/B testing technologies. Specifically, we look at the following tools: *ab tasty, adobe target standard, avenseo, beampulse, bunting, changeagain, cox'nductrics, con-*

vert, devatics, dynamic yield, experiment.ly, google content experiments, google optimize 360, google website optimizer, iterable, kaizen platform, kameleoon, kiss-metrics, 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, site-gainer, sitespect, split optimizer, stetic, visual website optimizer, visual website optimizer, and *zarget.* The figure replicates Figure 2 in the paper but using this more expansive definition. If anything, the effects appear somewhat larger. If our narrower definition of A/B testing leads us to treat A/B testing as controls, and A/B testing improves performance, then moving to a more expansive measure would increase our estimates.

[Figure A1 about here.]

A3 Regression tables corresponding to our specification chart

Here we report the regression tables that correspond to our specification chart in the body of the paper.

[Table A2 about here.]

[Table A3 about here.]

[Table A4 about here.]

[Table A5 about here.]

A4 Example Product Launches

We measured a startup's cumulative number of product launches by counting the number of weeks where a news article linked to the startup includes the strings "launch" or "introduce" in the article title. We use the Crunchbase news API to pull articles linked to the funded startups in our sample. Table A6 includes selected examples showing how our simple keyword matching algorithm works and where it fails. The majority of articles that match to "launch" or "introduce" cover product launches, though there are false positives. For example, in line 11 there is an article about LaunchCode changing its leadership team. In row 15 there is an article about a startup bootcamp in Edmonton tagged to a startup that participated in the program. In row 21 there is an article about a company thinking about launching a product in the future.

Table A7 shows examples of articles that match to startups in our sample that are not tagged as covering product launches. For the most part, these articles do *not* appear to be covering product launches but instead reporting on fundraising rounds and general interest news. That said, there do appear to be some false negatives. Rows 8 and 12 are articles both cover new features and extensions to products. In row 8, the app added the ability to share to multiple social networks at once. In row 12, the app added direct messaging. Both could be considered new products or classified as more "minor" extensions to existing products. While there is not absolute definition the advantage of using the words "launch" and "introduce" as it captures product changes that the media (and likely the startup's PR lead) deemed important enough to be thought of as new and different enough to explicitly call out.

[Table A6 about here.]

[Table A7 about here.]

A5 Additional product and website code results

Here we report regression tables for the results presented in Figure 5 along with further results showing that A/B testing impacts product development and the discovery of product-market fit.

To further investigate how A/B testing impacts product development, we also measure the (logged) number of new lines added and the number of lines deleted to check if A/B testing firms are introducing more code or simply deleting inefficient parts of their website. Complementing our "Major Change" measure, we also calculate if A/B testing firms are more likely to make minor changes, making code change that is the bottom 5% of the code change distribution. Finally, we also use a variant of the Ratcliff/Obershelp Gestalt Pattern Matching algorithm to generate a dissimilarity score from 0 to 1 for how a website changes between t and t'. This algorithm recursively looks for long common sub-strings within a text, assigning higher similarity to the text where there is more overlap in the sub-string tree. It is similar to other distance metrics like Jaccard similarity and Levenstein distance. We use the implementation in Python 3, which removes commonly occurring sub-strings (e.g., empty lines). This serves as a robustness check of our other structural and style measures. We regress these measures, along with the measures reported in Figure 5, against whether a firm has adopted A/B testing tools in Table A9 (except our number of products launched measure which we report at the start of Table A9).

We also analyze an additional five measures of whether A/B testing drives a firm to more quickly discover product-market fit. First, we use data from SimilarWeb on the website's weekly bounce rate, pages per visit, and visitor duration. The bounce rate is the percentage of visitors who immediately leave the site. For pages per visit and duration, we log the measures so we can interpret effects as percentages. Each of these measures captures variation in whether a startup has

improved engagement with its customers/users or if it is far from product-market fit. Our final measure of product-market fit is the (logged) amount of funding the startup has raised.

Given that investors are more likely to invest and invest more, once startups have found product-market fit, this measures gives us an indicator if A/B testing drives learning about product-market fit. These results are reported in Table A8. We only have data for all these variables for 173 weeks of our panel (instead of 208). This is because this data was pulled earlier in our analysis process before our data subscriptions expired. For the bounce rate, pages per visit, and duration variables, we only have measures for a subset of the startups for which there is enough data for Similar to generate estimates. Further, as discussed in the paper's body, the product launch data only covers startups that had raised funding at the start of our panel since coverage of launches by unfunded startups is incredibly limited.

[Table A8 about here.]

[Table A9 about here.]

A6 First stage of instrumental variable regressions

Here we report the first stage (adoption) regressions used in Table 3. Model 1 in Table A10 regresses Google 360 adoption on whether the firm adopted GTM at least a year before the launch. The coefficient is positive and very significant. Model 2 includes the technology stack control. The estimate does not change. Model 3 includes GTM interacted with the number of days since launch to account for time-varying adoption effects. Finally, Model 4 estimates the adoption

equation just using the day-before launch and the last day of our panel. We again find a positive and significant effect.

[Table A10 about here.]

A7 Generalized Synthetic Control Fit

Does our generalized synthetic control model do a good job estimating startup growth? Here we report the actual and estimated growth trajectories for 6 of the startups that adopted A/B testing in our sample. Overall, the fit appears relatively tight. While the actual pageviews exhibit more variation on a week-to-week basis, the estimated growth rates match well before adoption. Further, the estimated trends post-adoption appear to be reasonable extrapolations from past data.

[Figure A2 about here.]

Table A1: While larger firms are more likely to use cloud font providers once we account for our technology stack control and firm-fixed the estimated effect goes to zero.

	(1)	(2)	(3)	(4)
	Log(Visits + 1)			
Using Cloud Font Tool?	1.496***	-0.141***	0.982***	-0.016
	(0.031)	(0.019)	(0.017)	(0.016)
Constant	5.075***	0.968***	5.432***	1.225***
	(0.028)	(0.044)	(0.012)	(0.039)
Observations	7,334,496	7,334,496	7,334,496	7,334,496
Week FE	Y	Y	Y	Y
Firm FE			Y	Y
Technology Stack Control		Y		Y

Standard errors clustered at the firm-level in parentheses

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

Table A2: Specification chart results for the full panel models with the log+1 transformed dependent variable.

	Log(Visits + 1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Using A/B tool?	0.131***	0.061**	0.066***	0.056**	0.047*	0.021*
	(0.022)	(0.019)	(0.011)	(0.021)	(0.019)	(0.008)
Observations	7,334,496	7,334,496	7,299,234	1,119,872	1,119,872	1,114,488
Week FE	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y	Y	Y
Startup Growth FE		Y			Y	
Lagged D.V.			Y			Y
Event Study Sample				Y	Y	Y

Robust standard errors clustered at the startup level in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3: Specification chart results for the full panel models with the inverse hyperbolic sine transformed dependent variable.

	IHS(Visits)					
	(1)	(2)	(3)	(4)	(5)	(6)
Using A/B tool?	0.123***	0.060**	0.054***	0.048*	0.048*	0.021*
	(0.023)	(0.020)	(0.012)	(0.020)	(0.020)	(0.009)
Observations	7,334,496	7,334,496	7,299,234	1,119,872	1,119,872	1,114,488
Week FE	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y	Y	Y
Startup Growth FE		Y			Y	
Lagged D.V.			Y			Y
Event Study Sample				Y	Y	Y

Robust standard errors clustered at the startup level in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Specification chart results for the full panel models with 0 page visits weeks dropped from the panel.

	Log(Visits + 1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Using A/B tool?	0.126***	0.049***	0.089***	0.050***	0.027*	0.021*
	(0.015)	(0.011)	(0.010)	(0.015)	(0.011)	(0.008)
Observations	5,848,601	5,848,601	5,814,784	1,004,782	1,004,782	1,114,488
Week FE	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y	Y	Y
Startup Growth FE		Y			Y	
Lagged D.V.			Y			Y
Event Study Sample				Y	Y	Y

Robust standard errors clustered at the startup level in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Specification chart results for the "2x2" difference-in-differences model.

	Log(Visits + 1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Using A/B tool?	0.121*	0.225***	0.323***	0.064	0.121*	0.191***
	(0.058)	(0.054)	(0.043)	(0.059)	(0.054)	(0.043)
Observations	70,524	70,524	70,524	10,768	10,768	10,768
Week FE	Y	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y	Y	Y
2X2 Length	1/2 Yr	1 Yr	4 Yr	1/2 Yr	1 Yr	4 Yr
Event Study Sample				Y	Y	Y

Robust standard errors clustered at the startup level in parentheses

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

Table A6: Examples of articles we tag as covering product launches.

Articles Title (match keyword highlighted in **bold**)
1) Identified Technologies **Launch**es IoT Truck Tracking Device to Move Earth Faster and Cheaper
2) Pogoseat Partners with the Utah Jazz, Moves into the Concert Space and Successfully **Launch**es VIP Upgrades
3) Price-Tracking Service Nifti Switches Gears, **Launch**es Social Polling App Cinch
4) Peepla **Launch**es Live Streaming Service
5) Instagram **Introduce**s Partner Program To Help Businesses Market Better On Instagram
6) Peepla **Launch**es Live Streaming Service
7) Online education for the pros: Udemy **launch**es corporate training tools
8) Startup MasterClass Raises $15 Million, **Launch**es Kevin Spacey Acting Tutorial
9) Want To Know How Much More You Could Earn? Adzuna **Launch**es Free Market Insight Tool For Jobseekers
10) MapD **Launch**es Lightning Fast GPU Database And Visual Analytics Platform; Lands $10M Series A Funding
11) This week in tech: Leadership change at **Launch**Code, Square expands in St. Louis
12) Postmates **launch**es delivery API, announces partnerships with Everlane, Threadflip, Betabrand, & more
13) Boostability **Launch**es BoostSocial 2.0
14) Uber **Launch**es Virtual Hackathon For API Developers
15) Startup Edmonton **Launch**es Tomorrow: 'We're Committed to Making Something Happen'
16) Slack **introduce**s a Google Drive bot and connection to Google Team Drives
17) Vancouver's Payfirma **launch**es mobile payment app for the iPhone (a Square for Canada)
18) Fleksy **launch**es public SDK, ups the ante on smart keyboards for iOS
19) Paytm **launch**es messaging product, Inbox. Will you use?
20) Hired, A Marketplace For Job Searchers, **Launch**es in New York
21) SoFi Said to be Exploring **Launch** of REIT as Range of Services Expand
22) Moprise Is **Launch**ing A Flipboard For The Enterprise
23) Stripe **launch**es Instant Payouts feature to all after pilot with Lyft
24) UberEATS, Instacart **launch** Tucson delivery services
25) Zomato **Introduce**s Full Stack Food-Tech Platform for Restaurants and Kitchens

Table A7: Examples of articles we tag as *not* covering product launches.

Articles Title
1) Product comparison service Versus gets social for better-weighted data
2) The Kaggle data science community is competing to improve airport security with AI
3) Zenefits' free software business model has been declared 'illegal' in the state of Washington
4) Dating app Wyldfire tries to avoid creeps by letting women take the lead
5) Dealstruck Closes $1.2M Seed Funding
6) DigitalOcean drafts Mesos to make its cloud more production-ready
7) There's something surprising going on with our current obsession with beards, according to one historian
8) Vine Now Lets You Share to Multiple Social Networks at Once
9) Medical App Figure 1 Raises $10 Million Series B to Boost Healthcare Knowledge
10) London Fashion: Meet 5 Startups Re-Shaping the Industry
11) Cloudflare Takes On AWS Lamdba at the Edge
12) Instagram takes on Snapchat with direct photo messaging
13) Currensee Receives $6,000,000 Series B Funding Round
14) Flexport CEO expresses some remorse in taking cash from Peter Thiel
15) Delivery Hero fuels up for acquisitions with fresh $49m round
16) 8 Asian Startups That Caught Our Eye This Week
17) Didi Kuaidi did more rides last year than Uber ever has
18) Flipboard's getting ads, courtesy of Conde Nast
19) Here's how Zenefits is trying to reinvent itself
20) Sunday Review: Grockit sells to Kaplan, Voxy raises $8.5m and Hoot.Me joins Civitas Learning
21) Wu-Tang Clan go to space for Impossible Foods and White Castle
22) HomeHero Locks Down $23M For Its Home Care Marketplace
23) In Race to Find Tech Talent, Piazza Opens College Homework Site to Recruiters
24) Outbrain And Gravity Failed To Comply With Privacy Rules, Watchdog Says
25) Drync wine app partners with retailers to offer pickup as option

Table A8: A/B testing leads to more code changes, especially the addition of new code and to larger changes. We find no evidence that A/B testing leads to less change in a website's style nor does it lead to more incremental code changes.

	(1) Log-Changes	(2) Log-Added	(3) Log-Deleted	(4) HTML	(5) CSS	(6) Sequence	(7) Big	(8) Small
Using A/B tool?	0.055*	0.079**	0.023	0.018**	0.016*	0.016**	0.013*	0.000
	(0.022)	(0.024)	(0.021)	(0.006)	(0.007)	(0.006)	(0.006)	(0.005)
Observations	93,048	93,048	93,048	93,048	93,048	93,048	93,048	93,048
Firm FE	Y	Y	Y	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y	Y	Y	Y
Months between Snapshots FE	Y	Y	Y	Y	Y	Y	Y	Y
Website Codebase Size Control	Y	Y	Y	Y	Y	Y	Y	Y
Technology Stack Control	Y	Y	Y	Y	Y	Y	Y	Y

Standard errors clustered at the firm-level in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Log-Changes is our logged+1 measure of the lines of code changes.
Log-Added is our logged+1 measure of the lines of code added.
Log-Added is our logged+1 measure of the lines of code deleted.
HTML is our measure for the relative change in the website's HTML tree.
CSS is our measure for the relative change in the website's CSS Style.
Sequence is our Ratcliff/Obershelp measure of sequence dissimilarity.
Big is a whether the code change is in the top 5% of changes.
Small is a whether the code change is in the bottom 5% of changes.

Table A9: A/B testing impacts a wide range of product-market fit and performance measures.

	(1) # of product launches	(2) Bounce Rate	(3) Log(Pages per Visit + 1)	(4) Log(Average Visit Duration + 1)	(5) Log(Total Funding Raised + 1)
Using A/B tool?	0.067**	-0.003	0.014**	0.042**	0.055*
	(0.026)	(0.002)	(0.005)	(0.013)	(0.027)
Observations	2,281,178	4,601,773	4,601,647	4,601,589	6,213,814
Week FE	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y
Tech. Stack Control	Y	Y	Y	Y	Y

Standard errors clustered at the firm-level in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A10: First-stage regressions showing that startups that adopted GTM at least a year before the launch of Google 360 are significantly more likely to adopt the Google 360 tool.

	(1)	(2)	(3)	(4)	(5)
	Adopts Google 360 A/B testing tool?				
Google Tag Manager (GTM)	0.018***	0.018***	0.007**	0.028***	0.027***
	(0.003)	(0.003)	(0.002)	(0.004)	(0.004)
Google Tag Manager (GTM) X Days from Launch			0.0002***		
			(0.000)		
Constant	0.005***	-0.044***	0.005***	0.009***	-0.059***
	(0.000)	(0.003)	(0.000)	(0.000)	(0.004)
Observations	4,359,264	4,359,264	4,359,264	41,916	41,916
Week FE	Y	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y	Y
Technology Stack Control		Y	Y		Y

Standard errors clustered at the firm-level in parentheses

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

Figure A1: The pattern of results in Figure 2 holds when we use a more expansive definition of A/B testing.

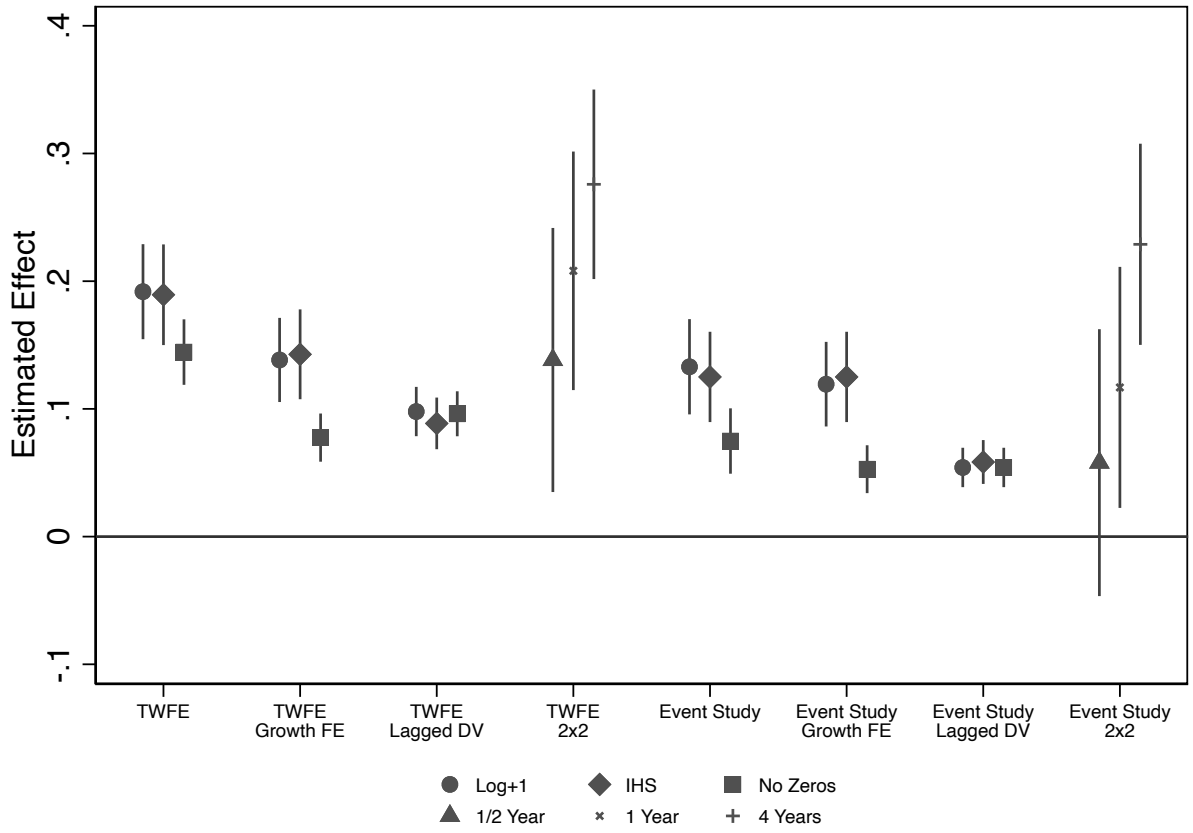


Figure A2: Actual and estimated growth trajectories for 6 firms that adopted Google 360. The gray vertical line is the date of adoption. The black line is the actual number of weekly visits. The dashed blue line is the estimate from our synthetic control model.