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

September 10, 2021

Abstract

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

1 Introduction

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

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

startup performance. A significant literature suggests that organizations have long learned from a variety of sources, including other firms (Mowery, Oxley and Silverman, 1996), outside experts (Cohen, Nelson and Walsh, 2002; Jeppesen and Lakhani, 2010), customers (Urban and Von Hippel, 1988; Chatterji and Fabrizio, 2014; Dahlander and Piezunka, 2014), suppliers (Dyer and Hatch, 2004), peers (Chatterji et al., 2019; Hasan and Koning, 2019), and even their own failures (Madsen and Desai, 2010). The learning process helps firms generate new solutions to their problems, assess their expected benefits, and select the most appropriate course to take. However, prior scholarship has also highlighted that the learning process is often biased (Denrell and March, 2001; Denrell, 2003): firms typically generate limited alternatives, rely on ad hoc processes to evaluate ideas, and make decisions based on practicality and instinct rather than data and analysis. Given these challenges to organizational learning, it is not clear whether merely reducing the cost of testing ideas will improve outcomes (Levinthal, 2017; Camuffo et al., 2020; Gans, Stern and Wu, 2019). For digital experimentation to matter, firms must also generate many alternatives to test *and* let the results of the A/B tests drive decision-making (Brynjolfsson and McElheran, 2016).

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

incremental search to reduce the risk of their most significant bets (e.g., a total redesign of a product) by modularizing a significant change into a sequence of smaller testable hypotheses.

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

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

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

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

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

We also contribute to a growing literature on the digitization of the economy. As the analytical capabilities of large firms increase, so does their productivity (Bryn-

jolfsson and McElheran, 2016; Brynjolfsson and McElherfan, 2019). In the spirit of recent studies that endorse an entrepreneurial process rooted in the scientific method (Camuffo et al., 2020), we demonstrate that startups also benefit from data-driven decision-making. Finally, our results echo a related marketing literature highlighting the importance of market testing for improving products (Urban and Katz, 1983; Boulding, Lee and Staelin, 1994; Runge and Nair, 2021).

2 Theoretical framework

2.1 Experimentation as an entrepreneurial strategy

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

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

While experimentation has long been promoted as a framework for strategic decision-making 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 between viable options is also challenging (Simon, 1959), while bureaucracy and other sources of inertia inside organizations may hinder the ability to take decisive action (Hannan, 1984). Finally, there are many, arguably less expensive, alternative channels 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.

However, in the last decade, rapid digitization of the global economy has altered this calculus (Brynjolfsson and McAfee, 2012). In particular, the cost of running controlled tests that compare alternatives has declined dramatically (Kohavi, 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., 2020). 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 published 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, and content to increase user engagement (Bapna et al., 2016; Kumar and Tan, 2015; Aral and Walker, 2014, 2011). Media companies A/B test the placement of articles or videos on their website, title variants, and subscription prices (Lawrence et al., 2018; Gomez-Uribe and Hunt, 2016).

2.2 A/B testing, learning, and performance

Nevertheless, organizational learning requires *more* than just a reduction in the cost of testing ideas (Fabijan et al., 2017)—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 investments in information technology yielded returns for firms only when they invested in hiring workers with relevant expertise (Brynjolfsson and Hitt, 2003). Likewise, the reduced cost of testing ideas may incentivize a firm to increase idea generation. John Cline, Director of Engineering at Blue Apron, highlights how an A/B testing platform led to more product ideas:[1]

> *Now that we have this capability, other groups have started using it. We went from one or two teams doing one or two tests a quarter to now, when we probably have at least 10 tests live at any given moment and a large number of tests every quarter being run by every product team.*

Another way that A/B testing supports idea generation is by reducing the impact of failed ideas and improving execution. For example, Emily Dresner, the CTO of

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

Upside Travel, notes:[2]

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

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

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

2.3 The alternative to formal experimentation

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

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

Next, prior work 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 led to productivity gains (Arrow, 1962). In our sample of high-technology startups, firms that are not conducting formal experiments may be tweaking their products informally, which may well lead to learning and improved performance, albeit at a slower pace. A/B testing should reduce the false positives of confirmatory search and accelerate the rate of discovering product improvements compared to tweaking. If, however, tweaking does lead to sustained gains in performance, A/B testing might have only a muted effect on firm performance.

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

3 Data and methods

To test whether A/B testing improves startup performance, 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 their homepage code over time from the Internet Archive's Wayback Machine to measure the degree and nature of change associated with adopting A/B testing. Appendix Section A1 describes each data source in detail.

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

After these exclusions, our primary 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 5, 2015, until March 24, 2019—amounting to 208 weeks (four years)—and a total of 7,334,496 firm-week observations.

We organize our analysis of this rich dataset into two distinct parts. First, we study whether A/B testing impacts startup performance through a series of sequentially developed models. After presenting these results, we use abductive reasoning and a second set of analyses to explore how A/B testing impacts startup performance.

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

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

This later part of our analysis, we iterated between theorizing, data analysis and data collection to arrive at most likely explanation for why we observe that A/B testing improves startup performance.

4 Does A/B testing impact startup performance?

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

4.1 Variable construction

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

Table 1 shows that just under 18% of firms use an A/B testing tool in the 208

[5]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 that 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.

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 slightly more than 20% of the market. The remaining 20% is split between Visual Website Optimizer, Adobe, AB Tasty, and Experiment.ly.

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

[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 five 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 zero, we add 1 before transforming the variable.

4.2 Two-way fixed effects performance effects

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

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

,

where Y_{it} is logged visits and our interest is in β, the impact of A/B testing adoption on startup performance. 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 as well as a host of other time-varying factors that could bias our estimates. Second, we include firm fixed effects α_i to control for time-invariant differences between firms. These factors could consist of the quality of the initial startup idea, the existence of a specific strategy, location advantages, and founders' educational backgrounds, 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).

Table 2 Model 1 estimates the raw correlation between A/B testing and weekly visits after accounting for week fixed effects. We find that firms that use A/B testing have 296% more visits than those that do not. In Model 2, we include technology stack controls. Suppose the adoption of A/B testing is correlated with the adoption and use of other technologies (e.g., payment processing tools). In that case, the raw estimate might reflect the impact of different technologies and not of A/B testing or major technological pivots that lead to better performance. The estimate drops to 19%, but the result remains precisely estimated and significant at the 0.1% level. Controlling for

time-varying technology adoption captures a meaningful amount of firm heterogeneity. In Model 3, we account for firm-level heterogeneity by including firm fixed effects. The estimated impact of A/B testing drops to 55%. Finally, Model 4 includes both our technology stack control and firm fixed effects. The estimate remains statistically significant, with a magnitude of 13%. The point estimate and standard errors suggests that A/B testing improves startup performance for the firms in our data.

[Table 2 about here.]

4.3 Alternative specifications and robustness checks

To further assess the robustness of our results, in Figure 1, we present estimates of the effect of A/B testing using a variety of different modeling choices.[6] The left-most estimate (the circle) presents the estimate from Table 2 Model 4. Moving right, the next estimate (the diamond) shows the estimates swapping out our logged-plus-one dependent variable for the inverse hyperbolic sine transformation. The choice of transformation does not alter the estimate. The third estimate (the square) excludes all observations 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 censored if they fail. The estimate remains unchanged. It does not appear that the overrepresentation of zero observations drives our findings.

[Figure 1 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 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 at the 1% level.

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

The estimates thus far control for variation in growth rates, but they mask time-varying differences in the impact of A/B testing on growth. For example, it could be that A/B testing has an immediate and small effect on growth, or perhaps its effect is gradual and compounds with time from adoption. Our TWFE results represent the average difference (within firms) between observations where A/B testing is used and firm-weeks where A/B testing is not. In essence, the TWFE can be thought of as an average of a multitude of "2-period" (e.g., week 1 vs. week 26, week 1 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 1, we show three estimates from models that include two periods of our data. The first (the triangle) comes from a simple difference-in-differences model using only data from the 1st and 26th weeks of our panel. As with all our models, it includes the technology stack control and firm and week fixed effects. The estimate is noisy, but the magnitude is similar to our baseline estimate at about 10% 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 last nine estimates in Figure 1 replicate the first nine but include only startups that transitioned to or from A/B testing.[7] These "event study" models rely exclusively on comparisons between firms that will adopt or have adopted. The estimates are somewhat smaller but still greater than zero and also suggest that the impact of A/B testing increases with time.

Using the event study specification also allows us to generate lead-lag plots showing the estimated effect size relative to when a firm starts and stops using A/B testing.

[7] 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.

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 2 about here.]

The top panel in Figure 2 provides estimates from a model that includes dummies for the 18 months before and after the switching event. The month before the event serves as the excluded baseline. Before a firm switches, there are few pre-trends, and the estimates overlap with zero. There are many reasons why we might expect a lack of observable pre-trends. If sales cycles drive adoption, it might lead firms to adopt A/B testing tools at points unrelated to other firm events. If an A/B testing provider significantly raises its tool's price, which firms like Optimizely have done in the past, firms may choose to stop using the tool for reasons unrelated to any 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.

The effect of A/B testing appears to grow with time, although the estimates in our event study graph are noisy. That said, they are consistent with our 2x2 models above, which more strongly suggests that the value of A/B testing increases over time.

The bottom panel in Figure 2 shows estimates for a model that includes dummies for 36 months before 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%, although the 95% confidence intervals are wide.

We report several additional robustness checks in the Appendix. In Section A3, we extend the robustness checks in Figure 1 by showing that changes to our definition of

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

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

4.4 Using the launch of Google Optimize to estimate instrumental variable and synthetic control models

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 remains. Second, our TWFE and event study estimates pool together many different tools and decisions to adopt and

dis-adopt. While this approach increases our statistical power and the generality of our results, causal inference remains a challenge because of the non-random adoption of A/B testing technologies.

To address these concerns, we focus on the global launch of Google's Optimize and Optimize 360 A/B testing suite on March 30, 2017. The tool, which integrates on top of Google Analytics, is similar to competitors 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 support. The Google Optimize 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.

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

To further test the robustness of these larger estimates, we also combine the Google

Optimize shock with synthetic control models to estimate the causal impact of A/B testing. The shock occurs two years into our panel, which allows us to leverage at least 100 weeks of pre-adoption data to construct matched synthetic controls for all Google Optimize adopters. Crucially, since no startup can adopt Google Optimize during this period, we can be confident that early Google Optimize adopters have not already selected out of our sample, nor have they adopted before we had enough observations to build a synthetic control. We use these synthetic controls to trace the expected growth trajectory had these startups not adopted the Google Optimize 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 of A/B testing appears to increase with time since adoption.

Specifically, we use generalized synthetic controls (Xu, 2017). This method has the advantage that it naturally averages across many different treated units, whereas the canonical synthetic control of Abadie, 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 3 shows the estimated treatment effect relative to the time of Google Optimize adoption. Panel A focuses on the effect trajectory a year before and after adoption; Panel B shows the complete set of estimates 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.[8]

[Figure 3 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

[8]In A9, we show the growth trajectory of six startups and the estimated counterfactual for that startup. The model appears to capture varied growth trajectories at an appropriate level of detail.

impact grows, and by six months out, the effect is significant. The plot suggests that the impact of A/B testing increases with time since adoption, although Panel B indicates that it stabilizes a little after one year.

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

[Table 3 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 that experience more substantial treatment effects. The TWFE models rely on a large panel that includes firms that adopt A/B testing for long periods and firms that install and quickly uninstall the technology. If the impact of A/B testing takes time to materialize and the TWFE includes more short-term adopters, this would lead to smaller estimates of the treatment effect. Suppose the IV compliers are more likely to adopt the tool for the long term, for example. In that case, because it integrates relatively seamlessly with the GTM tool they already rely on, we should expect the IV estimates to be larger. Overall, our findings suggest that researchers need to be aware that treatment effects may take time to materialize when analyzing the impact of experimentation practices.

In summary, these findings suggest that for the firms in our data A/B testing

improves startup performance and increasingly so with time. Moreover, A/B testing appears to enhance the performance of a wide array of startups, including those inside and outside of entrepreneurial hubs and in many industries, ranging from e-commerce to education.

5 How does A/B testing impact startup performance?

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

5.1 Can A/B tests lead to significant product changes?

Practitioners suggest that A/B testing is used for more than incremental experimentation (e.g., button colors or landing page copy). Instead, managers and engineers routinely note how experimentation has become a core part of their product development and deployment strategies. For example, engineers at Auth0 recommend an A/B testing strategy where startups "go big, then go small," lest they get stuck in a local maximum:[9]

> *When looking to A/B test, it's better to make big changes first to have the most significant effect. Small changes are good for optimizing further down the road, but those first need to be guided by larger-scale tests.*
>
> *In any industry, starting high-level and testing the entire landing page can open many doors for things to A/B test further down the line. Making big*

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

> *changes can reveal surprising and maybe even unintuitive results that you might've never expected.*

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

In addition, A/B testing may make the costs and benefits of innovation more transparent, helping managers build a robust quantitative case for a strategic shift. Joel Lewenstein, a product designer at the data collaboration software 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.[10]

> *Even the best qualitative theory rarely includes predictions about degree of impact. Testing and confirming improvements results in an outcome that can be weighed against other potential changes, and balanced against the work necessary to build and support that change.*

Further, aside from quantitative data, the existence of A/B testing tools provides further confidence that firms can learn in new or uncertain environments. Consider a company contemplating a market entry decision. If company executives are confident that they can quickly learn about customers in their new market and adapt using A/B testing, they may be more likely to enter in the first place. By accelerating learning after big strategic decisions, A/B testing can enable significant changes in the first place. Xu (2015) notes that quantification through A/B helps 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 practitioners have suggested that the real benefit of experimentation through A/B testing is the ability to conduct clearly defined experiments on

[10] 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

major changes to a product. EJ Lawless, a cofounder of Experiment Engine[11], reflects on thousands of A/B tests run on the company's platform:[12]

> *We also wanted to explore whether incremental changes (like Google's test of 41 shades of blue) or more radical overhauls tend to yield more successful tests. Clearly, radical changes tend to be more successful. This may be because they are more likely to have a clear hypothesis behind them, while incremental changes are more speculative.*

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

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

Finally, these insights also shed light on why relatively few startups adopt A/B testing tools. The quotes above show that effective usage of A/B testing tools requires more than installing a few lines of code and randomizing the color of a button. In-

[11] Acquired by Optimizely
[12] https://priceonomics.com/optimizing-the-internet-what-kind-of-ab-testing/
[13] https://www.optimizely.com/customers/hp/
[14] https://eng.uber.com/xp/

stead, when startups use A/B testing tools, they appear to shift how they generate ideas and decide between them. Similar to the arguments of Gans, Stern and Wu (2019), startups appear to make costly commitments in order to benefit from low-cost experiments. Indeed, the adoption results in Table 1 suggest that such investments are more challenging for smaller, non-funded early-stage teams to undertake compared to larger firms with more resources.

5.2 Measuring the relationship between A/B testing and significant product changes

To gain quantitative insight into whether firms in our data set are making significant shifts, we explore if A/B testing leads them to make more website and product changes. We collected from the Internet Archive's Wayback Machine to extract monthly snapshots of the code that generates each startup's home page. 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 vary in how much they change their code before the next snapshot in t'. We fit a model of the form:

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

,

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

is calculated at the monthly level. We also include a control for website size Log(Lines of Code) and fixed effects for the number of months between snapshots to account for larger websites, and longer durations between snapshots will feature more changes. However, our pattern of results does not shift if we exclude these two controls. We then calculate four different measures to evaluate changes to websites.

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

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

Relative change in HTML structure captures differences in the underlying code tree. We use an algorithm developed by Gowda and Mattmann (2016) that compares HTML tree structures returning a dissimilarity score that that ranges from zero (the same HTML tree) to 1 (completely different). This measure allows us to determine whether A/B testing is associated with small tweaks in copy or pricing or more significant changes to the structure of a website.

Relative change in CSS style uses a related algorithm that compares the similarity in the website's CSS to measure if A/B testing leads to incrementalism in design decisions. Again, this measure reflects a dissimilarity score ranging from zero (the same CSS style) to 1 (completely different). This measure allows us to test if A/B testing is more or less likely to change the visual presentation of a firm's product.

While the Wayback Machine allows us to test if A/B testing changes how firms

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

Products launched is calculated by parsing the titles in this set of Crunchbase-linked articles for each 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 A10 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 introducing new products. 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 4 shows how A/B testing impacts the startup's product development process.[15] The first result is the estimate on the number of lines changed between each Wayback Machine snapshot. This estimate suggests 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 fourth row shows that A/B testing increases the probability of a major code change. The final estimate suggests that A/B testing firms launch

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

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 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 4 about here.]

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

5.3 A/B testing, learning, and performance variability

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

A/B testing helps startups recognize which of the natural endpoints they are headed toward. Tests may reveal incontrovertible evidence that none of the startup's ideas are

high-quality. Moreover, the various changes startups make may not yield measurable performance gains on important metrics such as visits, clicks, or sales. Alternatively, experimentation could help a startup unearth a promising feature or customer segment that they can scale quickly (Azevedo et al., 2020). Armed with the data from a major A/B test, entrepreneurs can take decisive action about whether to persist or pivot to a more promising idea or company.

These dynamics suggest that while startups that use A/B testing should see increases in average performance, they should also be more likely to experience tail outcomes—scaling or failing.

5.4 Measuring if A/B testing increases performance variability

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

Figure 5 shows the estimated effect of A/B testing on whether a startup sample is more likely to scale or fail. Here, we see a bimodal response to the adoption of A/B testing just as predicted. This result suggests that A/B testing firms in our sample may be learning faster, 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 5 about here.]

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

6 Discussion and Conclusion

What is the right strategy for startups? Recent work in strategic management identifies experimentation as the preferred framework for decision-making in young businesses (Camuffo et al., 2020; 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 leads to a 10% increase in visits in the first few months after adoption for the startups in our sample. After a year of experimentation, the gains range from 30% to 100%. However, we find little evidence that the benefits of A/B testing vary between different kinds of startups. The most pronounced difference is between earlier- and later-stage startups, with early-stage startups appearing to benefit slightly more.

To explain these considerable performance improvements, we provide qualitative and quantitative evidence on how firms use A/B testing. Insights from practitioners suggest that A/B testing requires firms to shift their routines to realize the gains from experimentation. To complement this qualitative evidence, we demonstrate that firms using A/B testing launch more new products and make more significant changes to

their website code. Further, these firms are more likely to scale or fail, a performance pattern consistent with the A/B testing of consequential ideas.

Our article informs two research agendas at the intersection of strategy and entrepreneurship. First, while our field has generated many insights about strategy in large organizations, we have only recently sought to clarify entrepreneurial strategy. Our findings provide empirical evidence that an experimental approach to strategy, as suggested by Levinthal (2017); Gans, Stern and Wu (2019); Camuffo et al. (2020), leads to better performance for young firms. We offer a novel insight highlighting an essential distinction between running *an* experiment versus a strategy based on experiment*ation*. 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 experimentation will help firms quickly learn which of their ideas will significantly improve performance.

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

If experimentation is such a valuable framework for entrepreneurial strategy, however, why do fewer than one out of five firms in our sample ever adopt A/B testing? It could be that firms require complementary capabilities to leverage A/B testing fully. Indeed, we find that firms with more employees located in Silicon Valley and with VC

funding are more likely to adopt A/B testing tools. Further, much of the qualitative evidence referenced in our paper suggests that experimentation must be embraced by senior leaders and permeate the entire organization to impact performance. Thus, implementing this strategy in more developed organizations—with management teams, customers, and investors—is more complicated than in the case of lone entrepreneurs testing their minimum viable product (cf. Bennett and Chatterji, 2019; Camuffo et al., 2020; Felin et al., 2019).

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

Our approach is not without limitations. 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, although we take care to control important observed and unobserved factors that might drive A/B testing adoption and performance. Indeed, we show that our findings are robust when we use different identification strategies and methods ranging from instrumental variables to synthetic controls to difference-in-differences style models. In all our specifications, our results remain significant and consistent. Our effect sizes for technology startups, demonstrated on multiple metrics, are also in line with previous studies estimating the effect of data-driven decision making in large publicly traded firms (Brynjolfsson, Hitt and Kim, 2011).

Another challenge concerns generalizability. Our sample is composed of startups competing in digital markets, which can experiment at a low cost. However, beyond digital markets, the cost of experimentation likely varies widely across industries. These industry differences are helpful in comparing our conclusions to those of other recent

studies. For example, Gans, Stern and Wu (2019) describe an entrepreneurial strategy based on experimentation requiring some level of commitment that forecloses alternative strategic choices. The firms in our sample can likely experiment via A/B testing with fewer commitments than in manufacturing or life sciences, in which experimentation may be costly (cf. Pillai, Goldfarb and Kirsch, 2020). Future research can compare the impact of A/B testing on firms in these industries with other kinds of experimentation that require more significant trade-offs in the spirit of Gans, Stern and Wu (2019)'s "paradox of entrepreneurship."

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. Future research should investigate this phenomenon more deeply to better understand which organizational structures are most aligned with an experimental strategy. Finally, we do not observe the actual A/B tests that startups run, so our findings cannot discern whether A/B testing is part of a broader research and development program.

The continued decline in the cost of running digital experiments 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 run digital experiments, will widespread adoption alter the benefits to individual organizations? Finally, how will experimentation across the economy change the types of innovations that firms develop and how they are distributed (e.g., Kerr, Nanda and Rhodes-Kropf, 2014; Cao, Koning and Nanda, 2020)? Addressing these questions 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.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

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." *Review of Economic Studies* 29(3):155–173.

Azevedo, Eduardo M, Deng Alex, Jose Montiel Olea, Justin M Rao and E Glen Weyl. 2020. "A/B testing with fat tails." *Journal of Political Economy* 128(12).

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.

Bennett, Victor M and Aaron K Chatterji. 2019. "The entrepreneurial process: Evidence from a nationally representative survey." *Strategic Management Journal* .

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*. Digital Frontier Press.

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

Brynjolfsson, Erik and Kristina McElherfan. 2019. "Data in action: Data-driven decision making and predictive analytics in US manufacturing." *Rotman School of Management Working Paper* (3422397).

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 decision making 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. 2020. "A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial." *Management Science* 66(2):564–586.

Cao, Ruiqing, Rembrand Koning and Ramana Nanda. 2020. "Biased sampling of early users and the direction of startup innovation." *Harvard Business School Entrepreneurial Management Working Paper* (21-059).

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.

Deniz, Berk Can. 2020. "Experimentation and incrementalism: The impact of the adoption of A/B Testing." *Stanford GSB Working Paper* .

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.

Felin, T, A Gambardella, S Stern and T Zenger. 2019. "Lean startup and the business model: Experimentation revisited." *Long Range Planning* .

Flannery, Mark J and Kristine Watson Hankins. 2013. "Estimating dynamic panel models in corporate finance." *Journal of Corporate Finance* 19:1–19.

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

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

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

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.

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

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

Hubbard, Thomas N and Michael J Mazzeo. 2019. "When demand increases cause shakeouts." *American Economic Journal: Microeconomics* 11(4):216–49.

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.

Kerr, William R, Ramana Nanda and Matthew Rhodes-Kropf. 2014. "Entrepreneurship as experimentation." *Journal of Economic Perspectives* 28(3):25–48.

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).

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.

Luca, Michael and Max H Bazerman. 2021. *The power of experiments: Decision making in a data-driven world*. MIT Press.

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.".

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.

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

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." *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* .

Pillai, Sandeep D, Brent Goldfarb and David A Kirsch. 2020. "The origins of firm strategy: Learning by economic experimentation and strategic pivots in the early automobile industry." *Strategic Management Journal* 41(3):369–399.

Porter, Michael E. 1980. *Competitive strategy: Techniques for analyzing industries and competitors*. Free Press.

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

Runge, Julian and Harikesh Nair. 2021. "Exploration in action: The role of randomized control trials in online demand generation." *Stanford GSB Working Paper* .

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

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: Panel regressions showing how the inclusion of firm fixed effects and time-varying controls for a firm's technology stack reduce the estimated A/B testing effect from nearly 300% to just over 13%.

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

Weekly data from 2015 to 2019 on 35,262 Crunchbase startups founded between 2008 and 2013.
Linear regressions with robust standard errors clustered at the firm level in parentheses.
Brackets show 95% confidence intervals.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Point estimates from a generalized synthetic control model show that startups that adopt the Google's Optimize A/B testing tool see more growth and that this growth increases with time.

	(1)	(2)	(3)	(4)
		Log(Visits + 1)		
	1 week before	26 weeks after	52 weeks after	Average
Using Google A/B tool?	-0.001	0.371*	1.282*	0.676*
	(0.037)	(0.159)	(0.378)	(0.191)
	[-0.063, 0.076]	[0.058, 0.648]	[0.391, 1.829]	[0.246, 0.930]
Observations	4,359,264	4,359,264	4,359,264	4,359,264
Adopting Firms	618	618	618	618
Week 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

Weekly data from 2015 to 2019 on 20,958 Crunchbase startups that have yet to adopt A/B testing at the time of the launch of Google 360 in March 2017. Estimates are from a single generalized synthetic control model where the number of unobserved factors was selected by a cross-validation procedure.

Standards errors and confidence intervals calculated using a non-parametric bootstrap with N=500.

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

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

Figure 1: The effect of A/B testing on log page visits holds across a range of model specifications. "TWFE" indicates standard two-way fixed effects models, "Growth FE" indicates the model includes firm growth fixed effects, and "2x2" indicates that the estimate is from is a simplified difference-in-differences model that includes only data from the first week in our panel and a single observation either a half-year, year, or four years later. "Event study" indicates that only A/B switchers are included in the data. "IHS" indicates that we use the inverse hyperbolic sine instead of logged-plus-one visits. "No Zeros" indicates that all weeks where page views are zero have been excluded from the data. All models include startup fixed effects, week fixed effects, and a control for the size of the startup's technology stack. Bars are 95% confidence intervals.

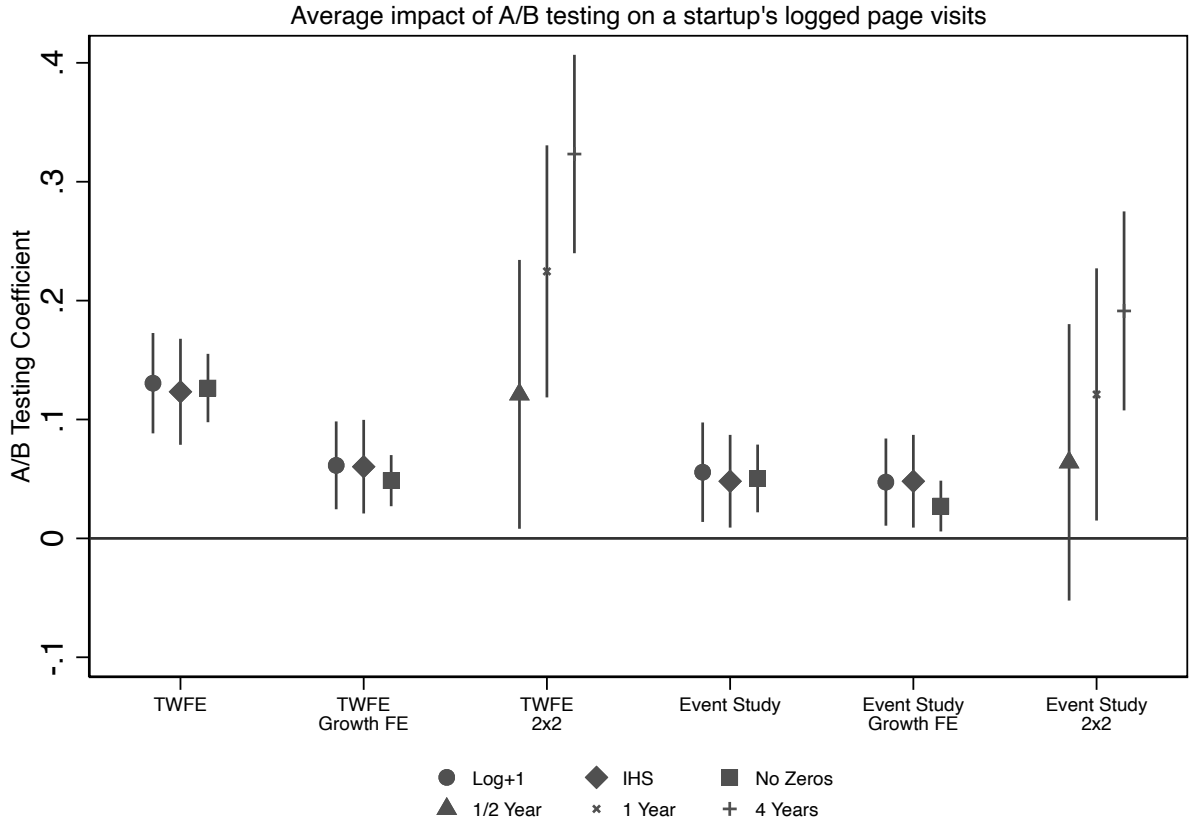


Figure 2: Event study plot showing the effect of A/B testing over time. Panel A shows the effect 18 months before and after use and Panel B shows a 36-month window before and after. The month before the event serves as the excluded baseline. The y-axis is the estimated coefficient for the A/B testing effect. All models include the technology stack control and account for week and firm fixed effects. The shaded region indicates 95% confidence intervals.

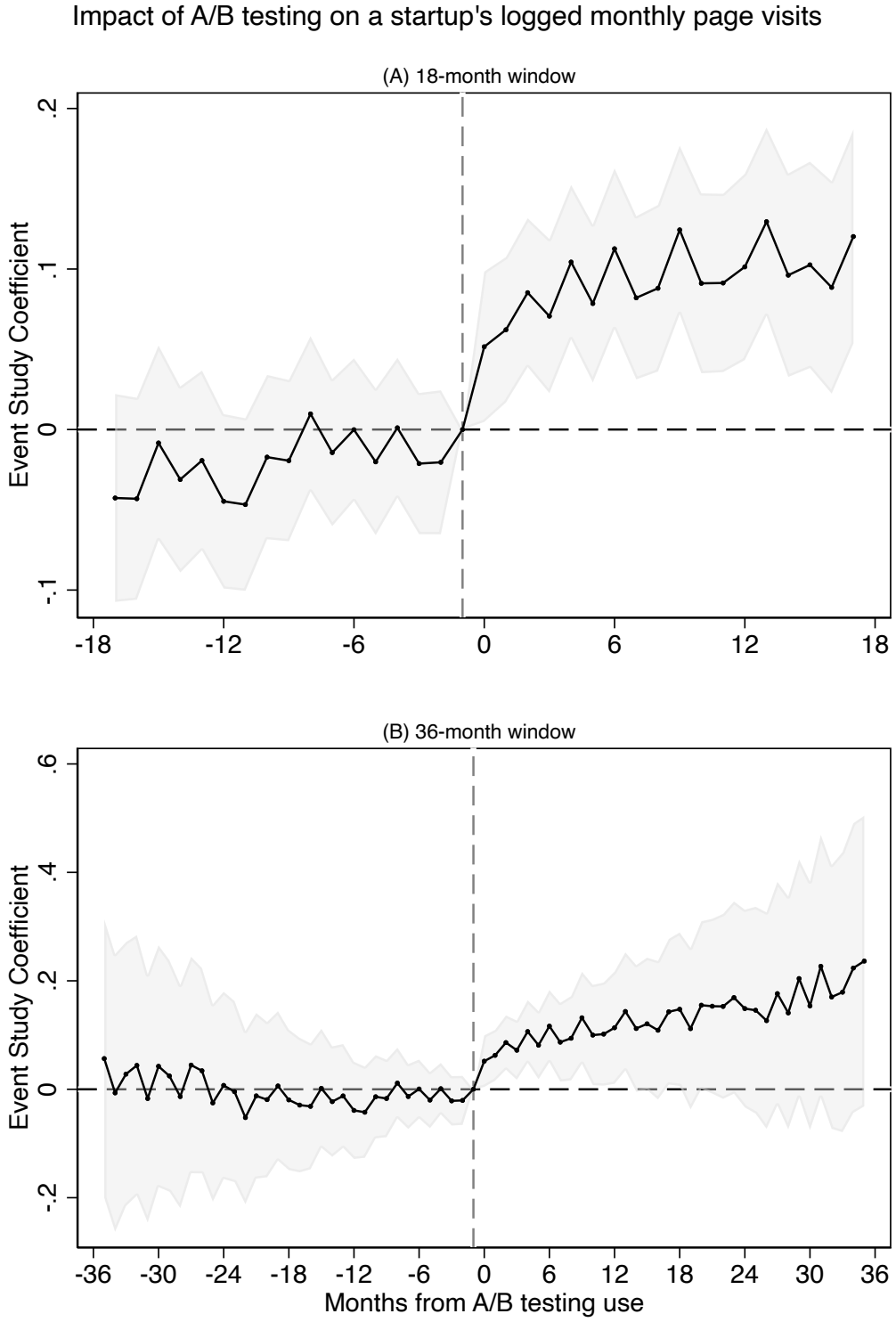


Figure 3: The estimated average treatment effect on the treated for the adoption of Google's Optimize A/B testing tool on logged page visits. Estimates are from a generalized synthetic control model. Panel A shows the estimated difference between an adopter and a synthetic control a year before and after the firm adopts. Panel B shows the same effects, but over a longer two-year pre- and post-window. The shaded region shows 95% confidence intervals.

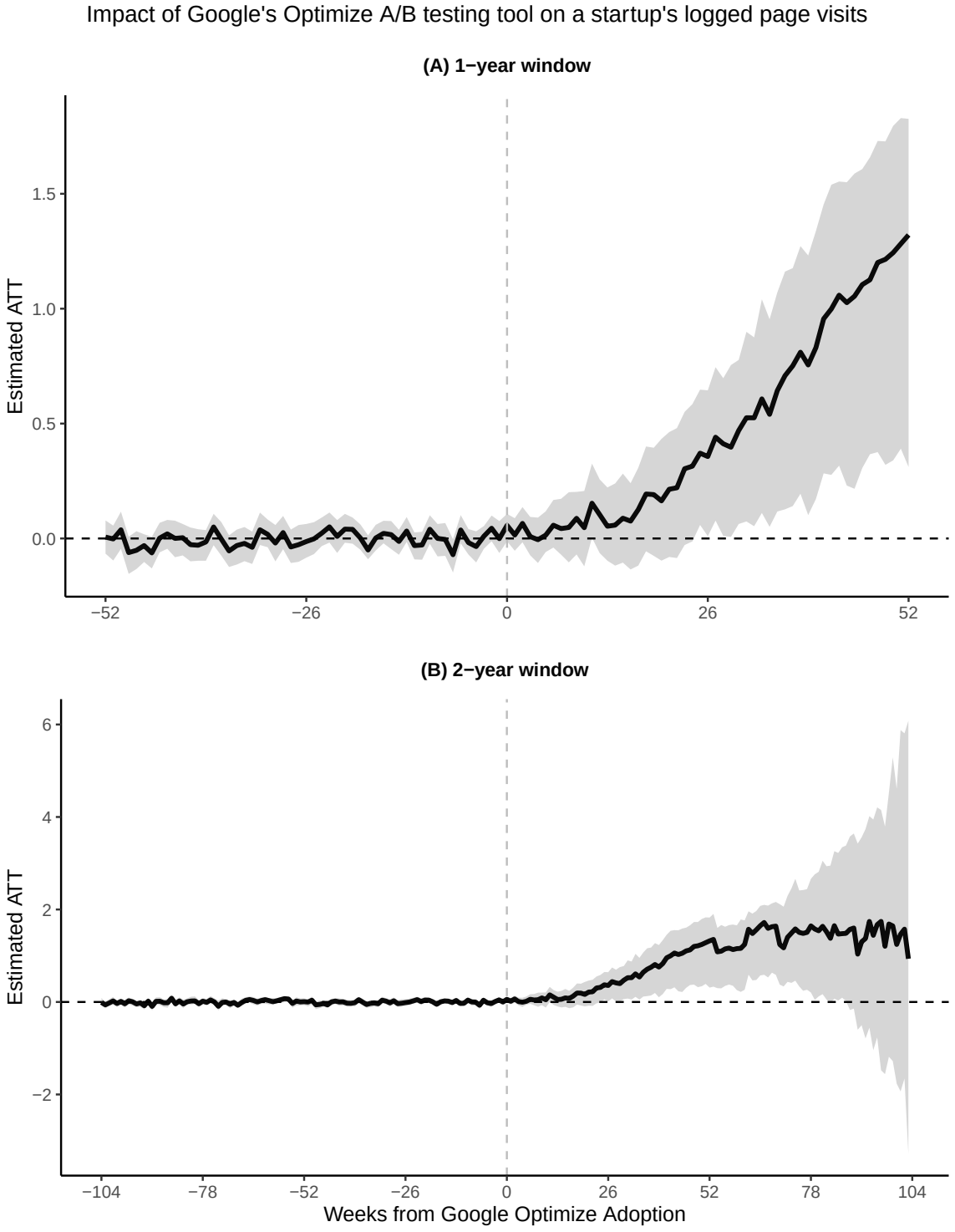


Figure 4: We find that A/B testing does not lead to incrementalism in product and website development for the nearly 10,000 startups for which we have website and product launch data. Instead, these firms make larger changes to their website code, the structure of their homepage's HTML, and website style and are more likely to deploy major code changes. A/B testing firms are also more likely to launch a new product in a given week than those that do not.

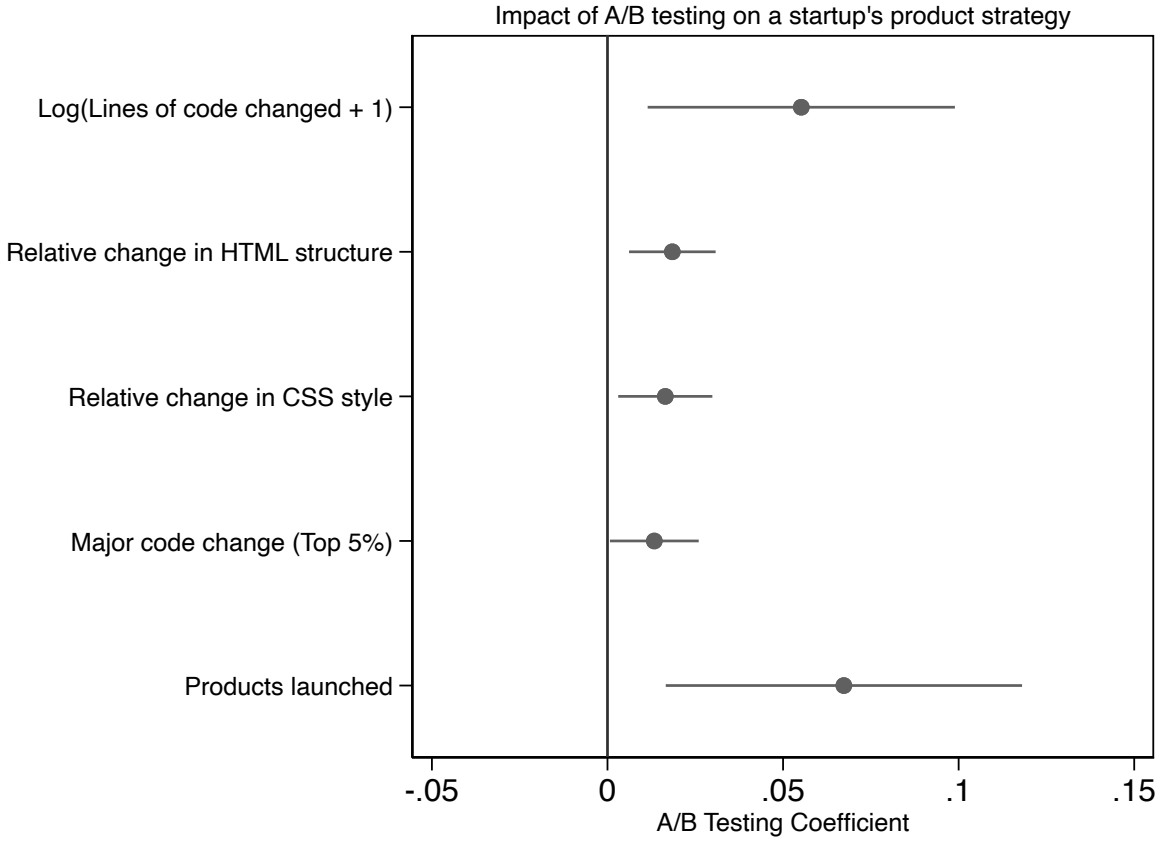
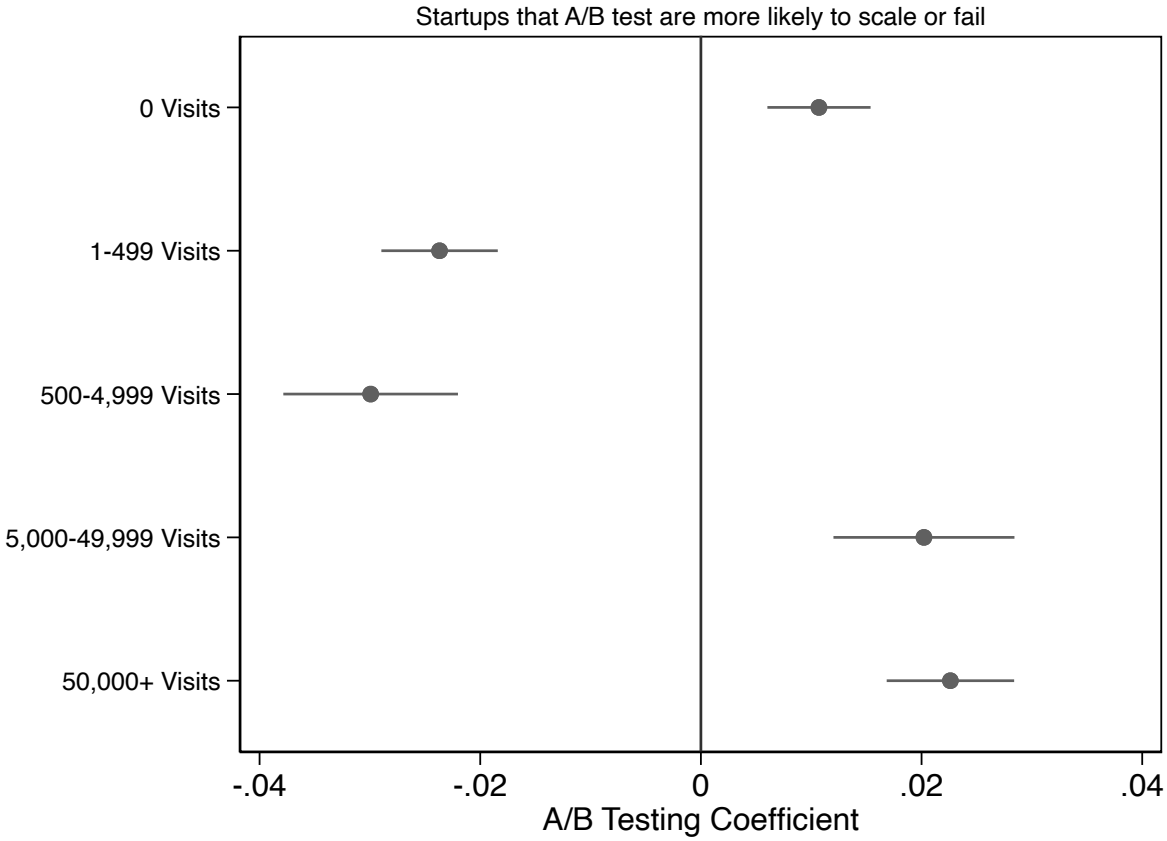


Figure 5: For the 35,262 Crunchbase startups in our primary sample, we find A/B testing firms are more likely to fail (end up with zero-visit weeks) and scale (achieve more than 50,000 visits in a week). The growth in tail outcomes comes at the expense of experiencing middling growth outcomes. Estimates are from a regression similar to Table 2 Model 4 but with the dependent variable dichotomized to reflect the visit ranges listed below.



Online Appendix:
Experimentation and startup performance

Table of contents

A1 Data sources

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. For each company. Crunchbase provides information including founding year, firm name, company website, funding raised, and a brief description of its product. The database also links to news articles covering investments, product launches, and other key company events. Finally, it also offers limited information on the founding team, executives, and board members.[16] This data is 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 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 over 250 million websites. It tracks these websites' current and prior technology stacks. Figure A1 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 for a website 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 A1 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

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

provides website performance metrics, including page views, bounce rates, and time-on-site over the last three years at the weekly level. SimilarWeb is used by firms like Airbnb, Procter & Gamble, and Deloitte for lead generation to track acquisition targets and 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 nonprofit 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.[17] 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.

A2 Regression tables corresponding to our specification chart

Here we report the regression tables that correspond to Figure 1 in the body of the paper.

[Table A1 about here.]

[Table A2 about here.]

[Table A3 about here.]

[Table A4 about here.]

[17] 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 the Wayback Machine very sparsely archived. The final API pull ran in February of 2019.

A3 Impact of A/B-enabled tools

In Figure A2, we show our findings hold across various 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, Conductrics, Convert, Devatics, Dynamic Yield, experiment.ly, Google Content Experiments, Google Optimize 360, Google Website Optimizer, Iterable, Kaizen Platform, Kameleoon, Kissmetrics, Leanplum, Marketizator, Maxymiser, Maxymizely, Mixpanel, Monetate, Myna, Omniture Adobe Test and Target, Optimization Robot, Optimizely, Optimost, Qubit Deliver, Roistat, Sentient Ascend, Shopalize, SigOpt, SiteGainer, SiteSpect, Split Optimizer, Stetic, Visual Website Optimizer,* and *Zarget.* The figure replicates Figure 1 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 A2 about here.]

A4 Placebo test using font libraries

In Table A5, we run a placebo test by checking if using cloud font providers increases growth. We have no reason to believe that adding a cloud font library would 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's font API, Font Awesome, fonts.com, MyFonts, Adobe Creative Cloud webfonts, Webtype, Mozilla fonts, Adobe Edge Web Fonts, and Font Plus.* 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 page views! However, as soon as we control for our technology stack, the effect turns negative. With firm fixed effects and our time-varying technology stack controls, our preferred model estimates that cloud font tools have a precisely estimated zero effect. The results in Model 3 highlight the importance of including firm fixed effects and accounting for time variation in the technologies a startup uses.

[Table A5 about here.]

A5 The effect of A/B testing on page visits is still positive when we model growth as a dynamic growth process.

In this section, we fit dynamic panel models that include the lagged dependent variable as a control. When it comes to causal inference, lagged models rely on a different conditional independence assumption than difference-in-differences type models (Angrist and Pischke, 2008). Further, by including the lagged variable, the interpretation of the coefficient on A/B testing changes (Flannery and Hankins, 2013). In the standard two-way fixed effect model, the coefficient on A/B testing represents the (weighted) average increase in page views post-adoption. In dynamic lagged models, the coefficient reflects the average increase in the weekly growth rate. In this case, A/B testing directly improves performance in the following week and future performance since the increased growth in the next week results in larger lagged growth two weeks later. Indeed, when reporting results from (AR(1)) dynamic panel models, researchers often also report the "long-run" effect by dividing the growth rate by 1 minus the lagged persistence term (e.g., Hubbard and Mazzeo, 2019).

Unfortunately, estimating dynamic panel models with firm fixed effects is non-trivial (Angrist and Pischke, 2008; Flannery and Hankins, 2013). Lagged variables end up correlated with current period errors, asymptotic approximations require both large T and N, and even commonly used instruments (like twice-lagged periods) rely on relatively heroic assumptions. Therefore, following Angrist and Pischke (2008), we fit a variety of different models to explore whether our findings hold when we model our data as a dynamic panel instead of using a difference-in-differences setup.

[Table A6 about here.]

Table A6 presents "pure" lagged dependent variable models in columns 1-3 and models with firm fixed effects and lagged dependent variables in columns 4-6. All models include our technology stack control and week fixed effects. Column 1 presents a simple model that includes a lagged dependent variable to control for growth dynamics. The model implies that A/B testing increases the weekly growth rate by 27% and that there is persistence in growth with a coefficient on the lagged term of of 0.84. The end result is sizeable long-term effect: 170%. Models 2 and 3 instrument the lagged term using twice-lagged and year-lagged dependent variables, respectively. The idea here is that if the process is AR(1), then prior errors should be uncorrelated with current period errors. In both models, the persistence term jumps to 0.97 and 0.99, suggesting that growth is quite sticky. While the A/B testing coefficients shrink to 3.8% and 0.3%, respectively, they remain significant at the 0.001 level. Further, since persistence increases so much, even a slight increase in growth can have a dramatic long-term impact. Indeed, both models suggest improvements in long-term growth from 165% to 300%.

Finally, models 4-6 replicate 1-3 but include firm fixed effects. The firm fixed effects reduce persistence and the main effect of A/B testing. However, the coefficient on A/B testing remains positive and statistically significant. While an order of magnitude smaller, the long-run effects range from 8% to 13%, consistent with the average post-adoption increase we find with our difference-in-differences style models. Overall, the results from these dynamic panel models support our findings in the body of the paper.

A6 Heterogeneous treatment effects

How does the impact of A/B testing vary across different startup types? In Figure A3, we present the estimates across different subsamples 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 subsample.

In Panel A of Figure A3, we analyze the heterogeneity in our effect based on startup characteristics—including factors such as venture funding, age, location, and 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, younger startups outside the Bay Area that are smaller and lack VC funding experience somewhat more significant effects from A/B testing. It is possible that early-stage startups that are not part of established entrepreneurial networks may learn faster using A/B testing than they would otherwise.

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 A3 about here.]

A7 Instrumental variable results

While the choice to adopt Google Optimize after it launches is endogenous, we construct an instrument for adoption by noting that Google Optimize strongly suggests firms also use Google 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 Optimize. 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 Optimize 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:

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

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

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

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 Optimize tool was launched. This scenario seems unlikely.[18]

We estimate our difference-in-differences instrumental variable model on the sample of startups at risk of adopting Google Optimize at the time of the launch. We do so by including only startups with the Google Analytics tool—a requirement for using Google Optimize and GTM—installed in the week before launch. Over 90% of startups in our data used 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 include only startups with more than zero page views at the Google Optimize launch in March 2017. By excluding these startups, we eliminate mostly 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.[19]

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

[19]There was an invite-only beta program for Google Optimize. In our sample, 23 firms are recorded as

We begin our instrumental variables analysis by checking that GTM use predicts whether a startup adopts Google Optimize. In Figure A4, we plot the percent of startups that have installed Google Optimize 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 Optimize, suggesting a robust first stage.

[Table A7 about here.]

We also report the first-stage (adoption) regressions in Table A7. Model 1 in Table A7 regresses Google Optimize 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.

[Figure A4 about here.]

Table A8 presents our second-stage IV estimates and the first-stage F-statistics. 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 A3 shows that the effect of A/B testing is more significant for smaller, younger, and less established firms. Suppose compliers are more likely to be these less established firms. In that case, it seems plausible that adopting Google Optimize could help a startup with 1,000 visits a week scale to have anywhere from 2,600 to 9,159 visits a week, the range of our 95% confidence intervals. While such magnitudes seem unlikely for a firm that already has hundreds of thousands of visitors, it might explain our estimates if the LATE excludes them.

[Table A8 about here.]

The second reason our IV estimates are substantial is that, while our first stage is strong (the F-statistic in our models are above 40), the absolute magnitude of

having Google Optimize 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 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 considers fixed effects and controls. Our first-stage estimate suggests that 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 Optimize, 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 Optimize 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. Appendix Figure A4 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%, although 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 identifies a LATE only 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 Optimize—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 Optimize is more effective when used with GTM; after all, Google recommends pairing these tools.

To address these concerns, we also estimate our effects using the fuzzy two-period 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 A4 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 Optimize 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 Optimize by our instrument during this period.

Column 3 in Table A8 shows the point estimate from the fuzzy DiD model, making the unrealistic assumption that the adoption rate in the non-GTM control group is stable. 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 columns 1-3 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.

A8 TWFE using Google Optimize

To further check the validity of our instrumental variable findings, we also estimate the effect of Google Optimize using standard TWFE models as in model 1 of Table 2 but with the sample used to estimate our instrumental variables. The estimate in Table A9 Model 1 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 estimated only 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 keep only GTM adopters that 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 A9 about here.]

Finally, in Model 4, we again utilize a 2×2 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 Optimize. 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.

A9 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 six of the startups that adopted A/B testing in our sample. Overall, the fit appears relatively tight. While the actual page views 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 A5 about here.]

A10 Example product launches

We measured a startup's cumulative number of product launches by counting the number of weeks when 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 A10 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, although 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 A11 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 that 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 no absolute definition of a product launch, the advantage of using the words "launch" and "introduce" is that we capture product changes that the media (and likely the startup's public relations lead) deemed important enough to call out explicitly.

[Table A10 about here.]

[Table A11 about here.]

A11 Additional product and website code results

Here we report regression tables for the results presented in Figure 4, along with additional 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 changes that are in 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 substrings within a text, assigning higher similarity to the text where there is more overlap in the substring 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 substrings (e.g., empty lines). This serves as a robustness check of our other website structure and style measures. We regress these measures, along with the website code measures reported in Figure 4, against whether a firm has adopted A/B testing tools in Table A12.

We also analyze an additional five measures of whether A/B testing drives a firm to discover product-market fit more quickly. We report at the start of Table A13 the regression model corresponding to the product launch results in Figure 4. The rest of the models in the table show alternative performance metrics. Columns 2-4 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 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 once startups have found product-market fit, this measure indicates if A/B testing drives learning about it. These results are reported in Table A13. 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 SimilarWeb to generate estimates. Further, as discussed in the body of the paper, the product launch data covers only startups that had raised funding at the start of our panel since coverage of launches by unfunded startups is incredibly limited.

[Table A12 about here.]

[Table A13 about here.]

Table A1: 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 A2: 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 A3: Specification chart results for the full panel models with zero 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 A4: 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 A5: While larger firms are more likely to use cloud font providers, once we account for our technology stack control and firm fixed effects, 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 A6: Dynamic panel models

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Visits + 1)					
Using A/B tool? [β]	0.273***	0.038***	0.003***	0.066***	0.016***	0.004*
	(0.007)	(0.001)	(0.001)	(0.011)	(0.003)	(0.002)
Lagged Log(Visits + 1) [λ]	0.839***	0.977***	0.999***	0.484***	0.870***	0.951***
	(0.001)	(0.000)	(0.000)	(0.002)	(0.001)	(0.009)
Long-run effect of A/B testing [$\beta/(1-\lambda)$]	170%	165%	300%	13%	12%	8%
Observations	7,299,234	7,263,972	5,500,872	7,299,234	7,263,972	5,500,872
Instrument for lag visits		2 Wk Lag	1 Yr Lag		2 Wk Lag	1 Yr Lag
Technology Stack Control	Y	Y	Y	Y	Y	Y
Week FE	Y	Y	Y	Y	Y	Y
Firm FE				Y	Y	Y

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

Table A7: First-stage regressions showing that startups that adopted GTM at least a year before the launch of Google Optimize are significantly more likely to adopt the Google Optimize 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

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

Table A8: Difference-in-differences instrumental variable models estimating the impact of the Google Optimize 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 A9: TWFE models estimating the impact of the Google Optimize 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 A10: Examples of articles our algorithm tags as covering product launches. The vast majority of articles appear to cover product launches, although there are a few false positives (see lines 11, 15, and 21).

Article Titles (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, uUps 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 A11: Examples of articles our algorithm tags as *not* covering product launches. The vast majority of articles appear to not cover product launches, although lines 8 and 12 look as though they might be false negatives.

Article Titles
1) Product Comparison Service Versus Gets Social for Better-Weighted Data
2) The Kaggle Data Science Community Is 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 Lat 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 A12: A/B testing leads to more code changes, especially the addition of new code and for more significant 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 A13: 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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A1: Example of the BuiltWith Data

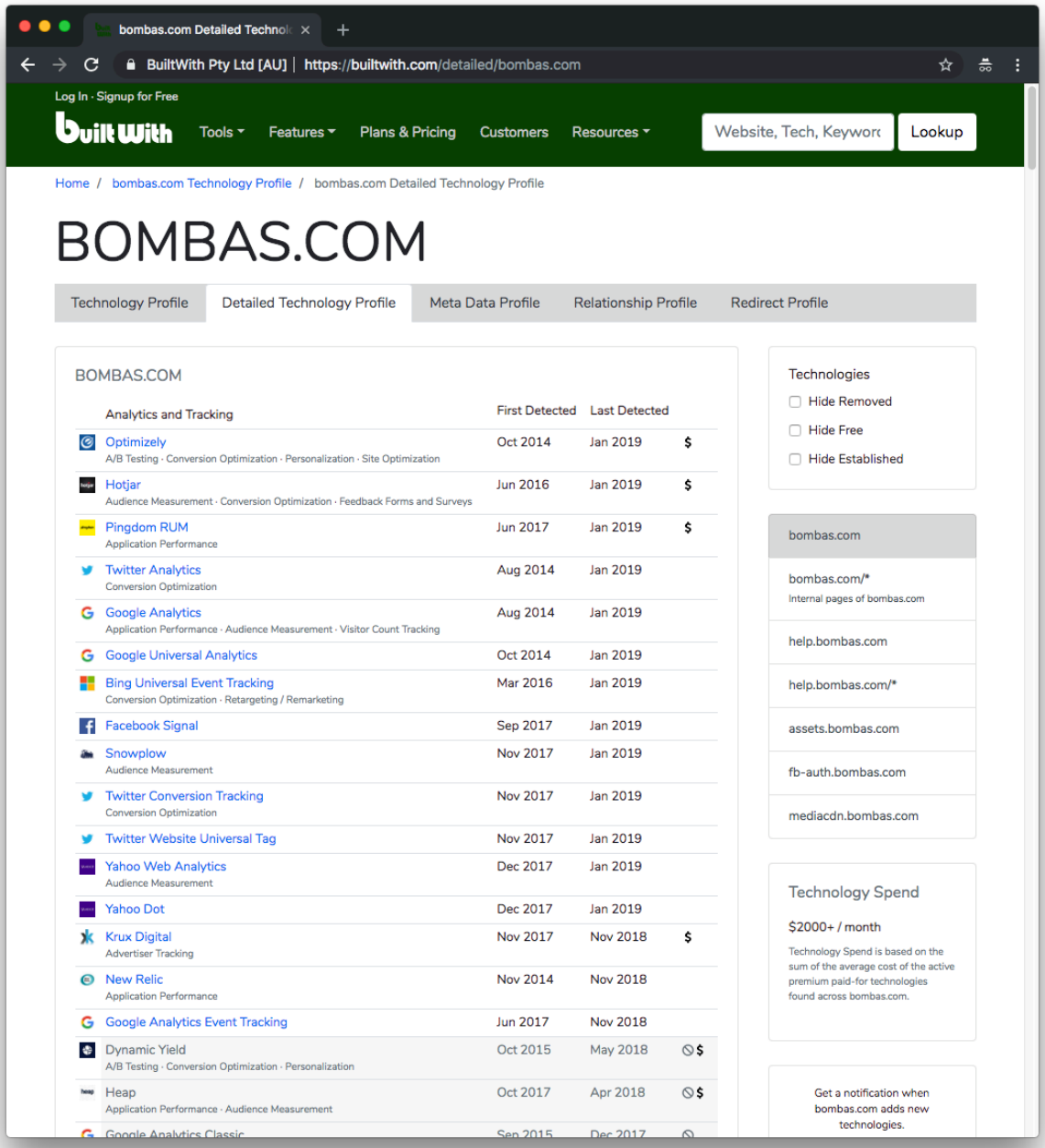


Figure A2: The pattern of results in Figure 1 holds when we use a more expansive definition of A/B testing.

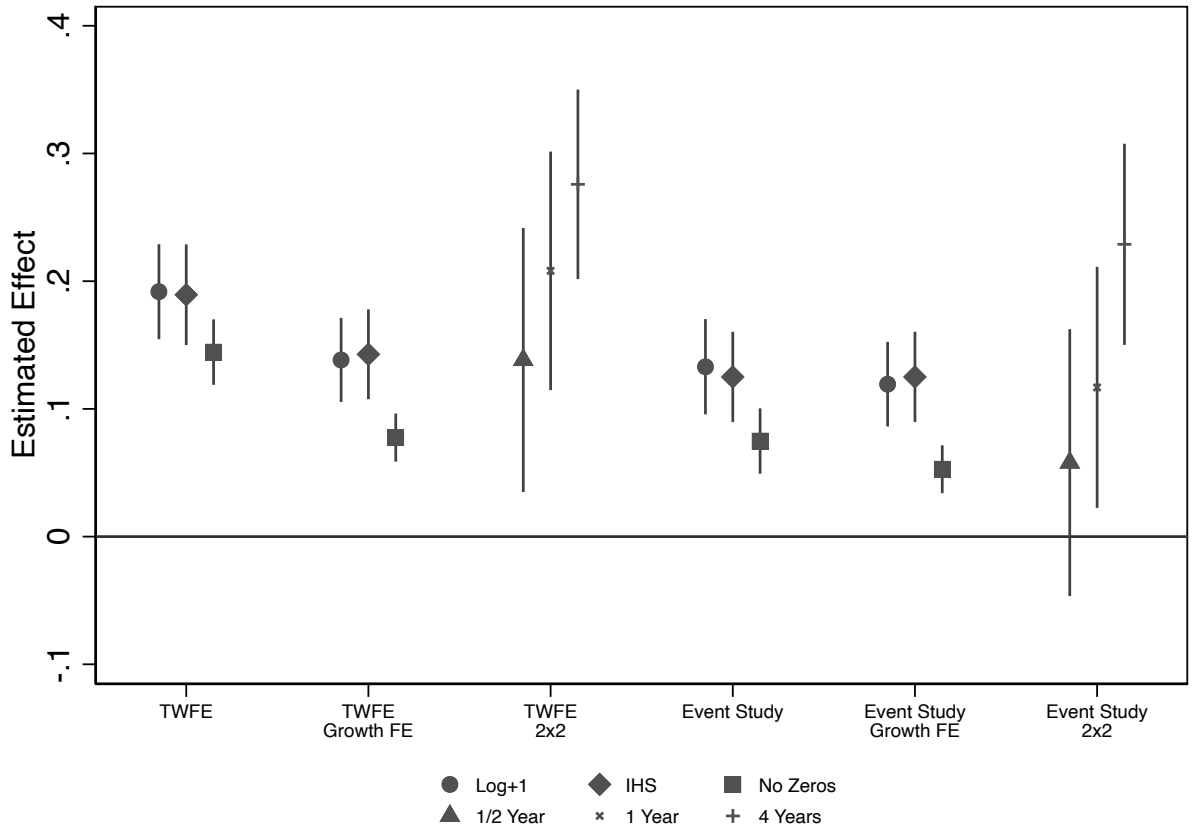


Figure A3: Heterogeneous effects of A/B testing by startup type (Panel A) and industry category (Panel B). Each estimate represents the effect of a A/B testing on growth for the sample of firms that meet the condition listed. Table 1 Panel B reports the number of firms of each type.

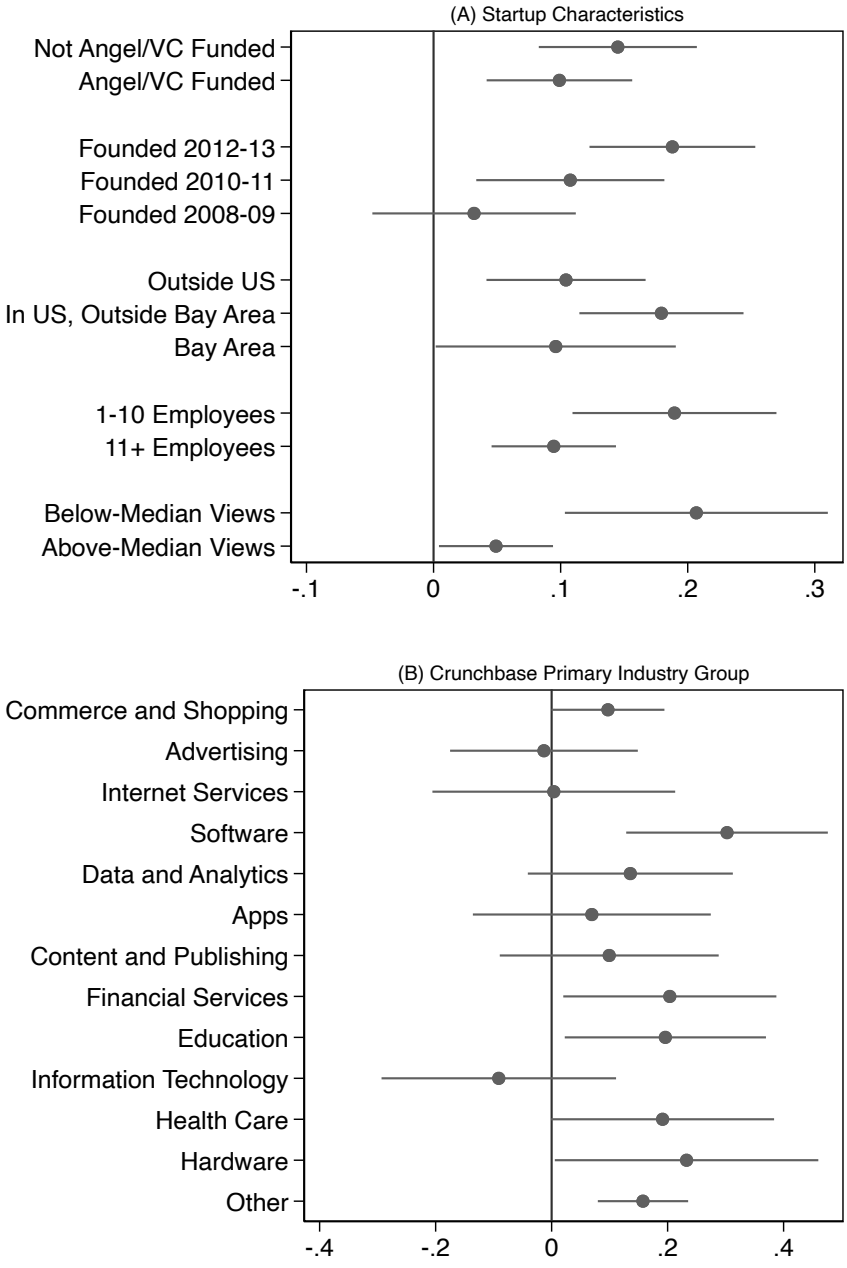


Figure A4: Adoption of Google Optimize by whether the startup adopted GTM at least a year before the Google Optimize launch. GTM startups adopt at significantly higher rates.

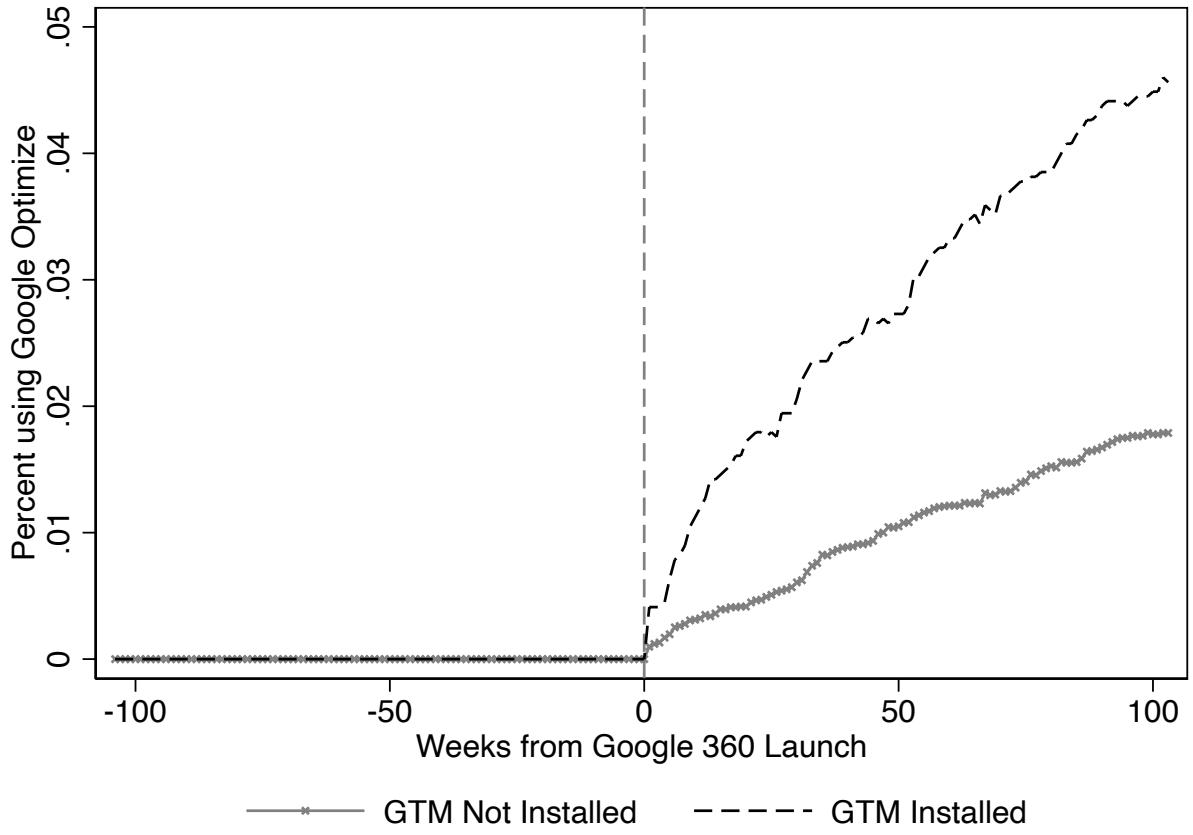


Figure A5: Actual and estimated growth trajectories for six firms that adopted Google Optimize. The gray vertical line is the date of adoption. The black line is the actual number of weekly visits. The dashed blue line is the estimate from our synthetic control model.