# Experimentation and Incrementalism: The Impact of the Adoption of A/B Testing

Berk Can Deniz

Stanford University, Graduate School of Business
Click Here for the Latest Version

October 10, 2020

## Abstract

This paper studies how the adoption of experimentation as a selection method shapes the direction of innovation. The spread of A/B testing, or digital randomized experiments, has made experimentation one of the most common methods organizations use to evaluate and select internally generated ideas. The strength of experimental evidence can be a force for radical innovation. By providing seemingly irrefutable evidence, well-designed and well-executed experiments can disabuse people of their false beliefs and generate breakthrough discoveries. However, I argue that the adoption of experimentation can also result in incrementalism, whereby firms focus on minuscule yet reliable improvements. These divergent outcomes can be explained by the incentives driving the people who design and implement experiments. The incentives of managers in established firms may lead them to use experiments in a way that undermines the pursuit of novelty while encouraging the search for incremental improvements. I investigate the relationship between experimentation and innovation in the context of US newspaper websites and their adoption of A/B testing. Using a historical archive of US newspaper websites and a novel computational method, I find that the adoption of A/B testing decreases the likelihood of radical change and makes websites more likely to change incrementally.

# 1 Introduction

> *Yes, it's true that a team at Google couldn't decide between two blues, so they're testing 41 shades between each blue to see which one performs better. I had a recent debate over whether a border should be 3, 4, or 5 pixels wide, and was asked to prove my case. I can't operate in an environment like that. I've grown tired of debating such minuscule design decisions. There are more exciting design problems in this world to tackle.*
>
> - Douglas Bowman (2009), former Design Lead at Google

Accurately evaluating and selecting ideas is a difficult yet crucial task for organizations and entrepreneurs that aim to innovate (Luo, Macher, and Wahlen 2020). Both potential breakthroughs and incremental improvements must be evaluated so that managers and entrepreneurs select ideas that have the most potential and allocate resources accordingly (Levinthal 2007). To gauge the underlying quality of ideas and predict their success, organizations employ various methods that range from in-house experts' judgment (Berg 2016; Keum and See 2017) to crowdsourcing (Csaszar 2018).

Even though these methods' ability to detect the highest-quality ideas has been frequently studied, what is less emphasized in the literature is the influence different selection methods have on innovation trajectories. The way in which new ideas are evaluated changes not only the type of ideas selected, but also the type of ideas generated in the first place. As a result, different selection methods have a potentially large influence on the degree of incrementalism and the balance between exploration and exploitation (Burgelman 1991; Knudsen and Levinthal 2007; Deniz and Sørensen 2019).

I investigate this dynamic in the context of randomized controlled trials. The recent popularization of A/B testing, or digital randomized controlled trials, has made experimentation one of the most common tools for evaluating and selecting new ideas. Thanks to third-party software platforms such as Optimizely, VWO, and Adobe Omniture, the cost of digital experimentation has declined sharply as it has become widely available for the arse-

nals of organizations and entrepreneurs (Kohavi and Thomke 2017). Using these platforms, even small organizations can conduct numerous experiments on their products and analyze results without hiring a single computer or data scientist (S. H. Thomke 2020b; Luca and Bazerman 2020).

The allure of experimentation as a selection method stems from the quality of the evidence it provides. A randomized controlled trial is widely regarded as the best tool for reliably measuring the difference between well-defined alternatives (Angrist and Pischke 2009), which is why randomized controlled trials are taught as the gold standard for scientific evidence. Compared to the use of observational data, surveys, anecdotes, and expert judgments, experiments produce rigorous evidence that is harder to dismiss (Kohavi and Thomke 2017).

The strength of experimental evidence can be a powerful force for innovation and discovery. A well-designed experiment can result in a fundamental breakthrough when it causes people to reject strongly held prior beliefs in the face of seemingly irrefutable evidence. For example, Galileo Galilei's famous falling bodies experiment refuted Aristotle's "the heavier an object, the faster it falls" thesis, which had been considered a well-established law of nature for almost two thousand years.

In the same way, business experimentation is seen as a force for innovative breakthroughs and a potential antidote to the exploitation bias in organizations (S. H. Thomke 2020b). Because the strength of experimental evidence can disabuse people of their false beliefs or faulty conjectures, it may push firms to be more exploratory and help them discover novel opportunities (Camuffo et al. 2019).

Nevertheless, experimentation does not necessarily result in exploration, radical innovation, or novelty. I argue that the ultimate effect of experimentation on the direction of innovation is often shaped by the incentives of the organizations and people designing and conducting experiments. Even though experiments have the potential to accelerate discovery and initiate breakthroughs, the incentives of managers in established firms (Manso 2011; Ederer and Manso 2013) often lead them to use experiments in a way that undermines the

pursuit of novelty while bolstering risk avoidance and the search for incremental improvements.

The idea that a focus on experimentation harms innovation is not uncommon among business practitioners. Brian Chesky, the CEO and co-founder of Airbnb, argues that "you can't A/B test your way to Shakespeare." Brian Norgard, the chief product officer of Tinder, argues that "A/B testing guiding product development is a clear sign that a company is out of new ideas." Similarly, some scholars have argued that relying on experiments as the sole method of evaluation can undermine exploration and result in incrementalism in a wide variety of settings, from management (Felin et al. 2019) to social science research (Akerlof 2020).

This paper studies how the adoption of randomized controlled experimentation as a selection method shapes the direction of innovation in established firms. I propose two potential mechanisms by which the extensive use of experimentation as a selection method exacerbates the bias for incrementalism and impedes exploratory search in established firms. First, the primacy of experimental evidence alters what can be evaluated and thereby selected. In other words, the primacy of experimentation can create a streetlight effect (Kaplan 2009), excluding novel ideas from the innovation process. Not every idea is equally feasible for randomized controlled trials, due to practical, financial, or ethical constraints. Furthermore, the feasibility of experiments is not random: Small, incremental improvements to the status quo are more suited to experiments than are novel ideas (Felin et al. 2019; Akerlof 2020). Hence, overfocusing on experimental evidence can undermine the search for novel ideas, advantaging the status quo and generating bias towards incremental ideas.

Second, the adoption of experimentation can alter which ideas managers choose to generate and test. A large body of literature has demonstrated that innovators in established firms are prone to act defensively, shun risky ideas, and test ideas that are more certain to succeed even when the risk to the organization is low (Ganz 2020; Aghion and Jackson 2016). This problem is aggravated by digital experimentation, which endows firms with the

ability to construct samples of unprecedented scale, generating enormous statistical power in experiments. As a result, managers can reliably measure and optimize even the most minute details of their websites, such as the shade of color for a button or the exact number of pixels in a border (S. H. Thomke 2020b; Kohavi, Tang, and Xu 2020). This novel ability to optimize the finer details of products and services exacerbates established organizations' existing bias towards exploitation and incremental improvements.

Hence, I predict that the adoption of experimentation as a selection method by established firms results in the erosion of the search for novel opportunities while boosting risk avoidance and incrementalism. I investigate the relationship between experimentation and incremental change in the context of the adoption of A/B testing by US newspaper websites. The empirical analysis is based on two novel data sets. First, using the Internet Archive's Wayback Machine, I constructed the historical archives of 297 active daily US newspapers' websites between 01 January 2014 and 01 January 2019. Second, I used a web intelligence firm's proprietary data set containing the historical records of different technological tools used by the newspaper websites, which allowed me to detect whether a website was using A/B testing technology and for how long. By combining the archives and the technology records, I analyzed US newspaper website design transformations and examined the relationship between those transformations and the adoption of A/B testing.

To observe and quantify website evolution, I employed a novel computational method that relies on the underlying HTML and CSS code of websites to measure the similarity between two different versions of the same website. This measure indicates how much change occurs on a webpage during a given period.

The results suggest an association between the use of A/B testing and incrementalism. After the adoption of A/B testing, the probability of large transformations decreases compared to the websites that are not A/B testing. In other words, A/B testing is associated with a slowdown in website transformation. This finding is valid for both on-average similarity over time and for the likelihood a website will be in the top decile of the largest

changes.

This paper contributes to the innovation literature and practice. The role of experimentation in innovation and entrepreneurship is growing rapidly as randomized controlled trials become a daily practice for many organizations. Despite a handful of recent studies, the academic literature lags significantly behind this real-world transformation. One of the most important questions yet to be answered about experimentation is how it relates to innovation. Seemingly contradictory arguments cause confusion about the role and potential of experimentation for innovation. Although we remain far from resolving the debate, this paper contributes to the discussion by presenting theoretical foundations. Moreover, this paper contributes to the literature on innovation by presenting a novel computational method and data source that can be used to examine innovation on websites, an area that has been neglected by the innovation literature due to data limitations.

The rest of the paper is organized as follows: Section 2 describes the conceptual model and potential mechanisms; Section 3 introduces the data sets, the methods used to mine those data sets, and the variables; Section 4 presents the results and examines alternative explanations; and Section 5 contains the discussion and the conclusion of the paper.

# 2 Theoretical Background and Framework

## 2.1 Selection and Evaluation

Even though multiple theoretical frameworks comprise the literature on innovation, all of them establish a clear distinction between the two stages of the innovation process: variation and selection (Nelson and Winter 1982; Burgelman 1991; Levinthal 2007; Knudsen and Levinthal 2007; Fleming and Mingo 2008; Girotra, Terwiesch, and Ulrich 2010; Berg 2016; Keum and See 2017). Regardless of the intended novelty of innovation, the process starts with the variation stage, in which organizations and entrepreneurs generate new ideas for products, services, or strategies. The amount of variation is important, as it increases the

probability of sampling ideas from the tails of the idea distribution (Girotra, Terwiesch, and Ulrich 2010; Azevedo et al. 2019).

The variation stage is followed by a selection stage in which new ideas' potential is assessed and compared to the status quo and to other ideas. Ideas with perceived potential are selected for further investment and implementation. Evaluating ideas and accurately calculating their underlying quality at this stage can be an extremely difficult task for firms and entrepreneurs. Given their limited resources, firms and entrepreneurs cannot afford to bring every idea to fruition and observe market performance. Consequently, ideas have to be evaluated and selected long before they are fully developed (Levinthal 2007). This necessity makes selection challenging; at the time of selection, there is still enormous uncertainty about an idea's future value.

The fundamental goal of organizations and entrepreneurs is to invest in ideas that will be successful and avoid committing resources to ideas that will not pay off in the future. Due to the difficulty of this task, organizations employ a diverse set of methods to gauge the quality of their ideas. Innovation scholars have studied many of these methods in detail, examining their accuracy and the biases that can distort them. The literature has overwhelmingly focused on the most common methods, such as managerial judgment (Reitzig and Sorenson 2013; Berg 2016; Keum and See 2017), domain expertise or experience (Scott, Shu, and Lubynsky 2020; Boudreau et al. 2016; Teplitskiy et al. 2019), and crowdsourcing (Mollick and Nanda 2016; Csaszar 2018; Luo, Macher, and Wahlen 2020).

What is less salient in the literature on evaluation and selection methods is the influence selection methods may indirectly have on innovation trajectories and outcomes. Special characteristics of selection methods, as well as the contexts in which they are used, might crucially influence innovation processes and their ultimate outcomes. For example, Knudsen and Levinthal (2007) examine how noise in the evaluation process and the level of hierarchy in organizations might influence the search process on the NK landscape. The authors demonstrate that perfect evaluation and hierarchy result in firms being locked on a local hill,

while imperfect evaluation and polyarchy result in a more robust search over the landscape. Moreover, Deniz and Sørensen (2019) examine how evaluation by a manager as opposed to a crowd interacts with idea generators' incentives. Deniz and Sørensen find that when idea generators know that an expert will evaluate them, they try to generate ideas that target the expert's taste, resulting in a decline of variation in their ideas. Hence, the way in which new ideas are evaluated changes not only the type of ideas selected, but also the type of ideas generated in the first place. As a result, different selection methods have a potentially large influence on the innovation process and outcomes.

## 2.2 A/B Testing

I focus on randomized controlled trials as a method for evaluating and selecting internally generated ideas. Prior to the internet, these trials took the form of physical experiments, which could be costly and time-consuming due to reasons ranging from the need to manufacture a physical prototype, to the process of gathering an adequate number of experiment subjects (S. Thomke 2008). Until recently, the cost of running a physical experiment prevented the majority of organizations from using experiments as frequently as they would have liked (S. H. Thomke 2020b).

Digitization has made business experimentation ubiquitous. The restrictions of the physical world, such as prototyping and sampling costs, are relaxed in digital environments (Kohavi, Tang, and Xu 2020). As a result, digital randomized controlled trials, or A/B tests, became extremely inexpensive to conduct and scale. Nowadays, as long as there is enough website traffic, firms can run and analyze multiple A/B tests every day without hiring a single computer or data scientist. This unprecedented access to randomized controlled trials is what I term the "democratization of experimentation." A simple demonstration of this trend can be seen in figure 1, which represents the increasing popularity of the term "A/B testing" in the Google search engine.

—INSERT FIGURE 1 ABOUT HERE—

## 2.3 Experimentation as a Selection Method

Experimentation as a selection method is appealing because of the high-quality evidence it produces. When it comes to measuring the differences between well-defined alternatives, a randomized controlled trial is widely seen as the most accurate method. The strength of randomized controlled trials arises from their internal logic and a design that aims to isolate the effects of a small number of variables with high certainty and confidence. Under certain assumptions, randomized assignment combined with a control group eliminate the fundamental problems of causal inference (Angrist and Pischke 2009; Kohavi and Thomke 2017).

Observational methods such as machine learning are often seen as alternatives to experimentation (S. H. Thomke 2020b). Even though observational methods can be used for evaluation and selection, causal identification constitutes a crucial distinction. Through randomization, experiments resolve the fundamental problem of causal inference without requiring strong assumptions or expertise of the context. When conducted properly, experiments generate the best evidence for the existence and size of differences between experimental conditions.

The strength of experimental evidence can be a force for innovation, swaying even the most strongly held beliefs. Because experimental evidence is harder to dismiss out of hand than observational or anecdotal evidence is, it can constitute a much tougher challenge to existing ideas. The history of science is full of examples of experiments that triggered fundamental breakthroughs in people's beliefs. In perhaps one of the most important randomized controlled trials in the history of medicine, Joseph Lister demonstrated that a large proportion of hospital infections could be prevented with simple sterilization practices (Howard 2013). Before Lister, post-surgical infection was a common cause of death. At the time, basic sterilization methods such as hand washing were not widespread; on the contrary, surgeons took pride in walking from one surgery to the next with their hands bloody. Lister, pointing toward early microbiology studies of Pasteur, argued that one potential reason for

the frequency of infections was the transfer of bacteria through blood. He argued that hand washing and other sterilization methods could prevent the transmission of bacteria from one patient to the other, thereby preventing infections. His claims were initially dismissed because the medical establishment had other theories about the cause of post-surgical infections. Yet, Lister conducted multiple experiments and demonstrated that simple sterilization did achieve a large reduction in infection-related deaths. Lister's work launched a worldwide revolution of surgical and medical practices.

The same dynamic applies to business experiments. A well-designed, well-executed experiment generates seemingly irrefutable evidence that can transform the strongest beliefs, making experimentation an exceptional tool for breakthrough innovation. Most established firms and their leaders may be entrenched in strong theories about their products and customers. The strength of experimental evidence has the potential to break these beliefs and disabuse managers and entrepreneurs of their false theories. Camuffo et al. (2019), for example, trained entrepreneurs in experimentation and the scientific method and then observed that those entrepreneurs were more likely to pivot or to terminate their start-ups. Camuffo et al. interpret this outcome as evidence that entrepreneurs became better at evaluating their beliefs and abandoning their false beliefs. Furthermore, Koning, Hasan, and Chatterji's (2019) study of start-ups' implementation of A/B testing conclude that cheap and easy experimentation makes it easier to evaluate a large number of ideas. They argue that this may motivate firms to generate more ideas, improving the chances of innovation and discovery.

Even though experiments can be a force for breakthroughs and a way to motivate innovation and exploration in organizations, the use of experiments does not necessarily favor exploration. On the contrary, overreliance on experiments can undermine innovation and bolster the search for incremental improvements. I argue that these divergent consequences are explained by the incentives of the people who generate ideas and design experiments. The incentives of managers, combined with the special characteristics of randomized controlled

trials, shape the direction of innovation through two potential mechanisms that predict that experimentation will undermine innovation in established organizations.

## 2.4 The Constraints of Experimentation

> *In the past, we could spend six months designing something: architecting it, building it, testing it. Now [A/B testing] gives us the ability to make quicker, better decisions ... [P]eople have started to think experimentation as a first, second step to ideas: could we test this, could we run an experiment?*
>
> - Al Booley, Senior Product Manager at BBC

Relying on experiments as the sole acceptable method for collecting evidence can impede innovative breakthroughs simply because many ideas do not lend themselves to experimentation. Perhaps the most well-known example can be observed in the tobacco and lung cancer controversy. In the early 20$^{\text{th}}$ century, a rising cancer rate was one of the most significant public health issues. Many scientists argued that there was a link between tobacco and cancer, and they presented overwhelming evidence that the increasing number of cancer cases was indeed a result of tobacco consumption. Nevertheless, due to ethical and practical obstacles, their argument could not be tested with a randomized controlled trial. However, some overzealous proponents of randomized controlled trials, such as Ronald A. Fisher, refused to believe the large body of non-experimental evidence. For people like Fisher, any piece of evidence not gathered under the rigor of an experiment could and should be immediately discarded as inadequate. As a result of such arguments, the tobacco industry managed to exploit the lack of experimental evidence and generate numerous alternative explanations to exonerate tobacco, slowing scientific progress for decades.

It is important to note that Fisher was right: experiments do provide more rigorous evidence. However, practical constraints, financial limitations, and ethical concerns will necessarily exclude many kinds of studies from randomized controlled trials. Failing to consider this simple fact can paralyze progress and thwart exploration.

Adopting experiments, in business as much as any other field, introduces a duality: these experiments may spearhead breakthroughs, but they may also prove to be a formidable obstacle to innovation. Radical discoveries and rapid exploration can be accomplished when organizations use experiments to test well-established beliefs. However, if the allure of the strength of experimental evidence is so strong that experiments become the only acceptable way of collecting evidence, experimentation can undermine the pursuit of novelty. In other words, diverse approaches to evaluating ideas can be discarded (as not rigorous enough) when organizations rely too much on the sense of certainty provided by experiments.

Although this may seem like an innocuous attempt to be more rigorous in decision making, the level of certainty provided by experiments may come at a high cost. As demonstrated by the tobacco and lung cancer debacle, experiments may not be feasible for many consequential ideas and questions. Because some ideas cannot be tested with a randomized controlled trial, valuing experimental evidence over all other kinds of evidence can lead to the abandonment of such ideas.

An organization's ability to disabuse itself from false belief is restricted to the degree to which the organization relies on experimentation. To illustrate, if any firm operates in such a way that established beliefs and theories can only be challenged by experimental evidence, then that firm's ability to disabuse itself of false beliefs will be significantly restricted in any case where opposing ideas are ill-suited to experimentation. Even in the presence of a large body of non-experimental evidence, such firms would dismiss any novel initiative that goes against their established beliefs. Despite their good intentions, such organizations may generate the same dynamic that undermined scientific progress that would have established the link between tobacco and cancer earlier.

A direct outcome of the primacy of experimentation is the streetlight effect (Kaplan 2009), in which organizations prioritize ideas that are suitable for experiments while excluding from the innovation funnel any ideas that cannot be quickly or easily tested. This dynamic might constitute a significant obstacle for the pursuit of novelty in two ways.

First, by the virtue of being novel, new ideas are less well understood compared to the status quo. As such, they require time and effort to be carefully formulated and translated to a testable hypothesis. Incremental ideas, on the other hand, are already well understood given their proximity to the status quo. Translating them into experiments is more straightforward. Consequently, if experimentation is the first step in the evaluation of every idea, novel ideas will always be at a disadvantage (Felin et al. 2019; Akerlof 2020).

Secondly, given the bias for experimental evidence, organizations are more likely to discard "softer" forms of evidence such as observational data, surveys, or case studies, despite the tremendous amount of support they can provide for novel ideas that cannot be easily tested with an experiment. Discarding or devaluing this type of evidence also disadvantages novel ideas (Akerlof 2020).

Put simply: a reliance on experiments may sacrifice exploration in favor of securing reliable measures of the difference between well-understood alternatives. Established firms and managers would be more likely to embrace this trade-off, as they are incentivized towards safety and certainty. Experiments provide them with the opportunity to be more conservative in their decision-making than ever before.

## 2.5 Managers Likely to Test Incremental Ideas

Although the intended purpose of randomized controlled trials is to test new ideas, organizations sometimes use the outcomes of those trials to judge the competency of the people who design the experiments. This dynamic feeds a well-known agency problem within organizations: even when the principal wishes agents to take risks, agents tend to shun risky ideas and become inert in order to avoid failure and subsequent negative evaluations (Ganz 2020; Gibbons and Roberts 2012; Aghion and Jackson 2016).

This agency problem will, I argue, be aggravated when randomized controlled trials are the method of evaluation. As discussed above, randomized experiments produce higher-quality evidence with fewer biases and noise than do other types of evidence such as ob-

servational data or anecdotes. In other words, when an experiment demonstrates that an idea is better than an alternative, it is, in fact, more likely that the idea is indeed of higher quality; little room is left for alternative interpretations. Because experiments more reliably reveal the true quality of an idea, they can also be perceived to reveal the quality of the idea generator. Managers who generate and test ideas cannot find alternative explanations when their ideas fail. Consequently, experiment outcomes can be seen as an appropriate way to evaluate a manager's quality.

When the people who generate ideas and design experiments think, sometimes correctly, their career trajectories are dependent on the outcomes of experiments, the fear of being negatively evaluated might push them to be more conservative. This dynamic could lead managers to avoid exploratory ideas, which have a high failure rate. Instead, they would experiment with incremental, exploitative ideas that are minimally risky and more likely to have a positive impact.

In recognition of how incentives can work against innovation, most of the practitioner-oriented literature emphasizes the importance of organizational culture for experimentation (Kohavi and Thomke 2017; S. H. Thomke 2020a). Practitioners are told to embrace failure as a given and transition away from traditional management practices that might punish such results. The existence of this discourse implies that some organizations do use experiment outcomes to evaluate and punish managers. If the agency problem is indeed prevalent among practitioners, then it constitutes a serious force that can impede innovation efforts within organizations.

### 2.5.1 Large Statistical Power and the Optimization of Minute Details

> *I do associate A/B testing with a kind of intellectual laziness that can often be harmful and cause product teams to lose sight, but the benefits are so large that it's not like we need to throw the baby out with the bathwater.*
>
> - Sean J. Taylor, Research Scientist at Lyft

A/B testing exacerbates the problem of managers tending toward incremental ideas for experimentation. A special feature of A/B testing is the ease and low cost of collecting gigantic samples. Websites can direct their visitors to experiments without offering payment or seeking explicit consent. As a result, it is not uncommon for A/B tests to have sample sizes in the millions (Kohavi and Thomke 2017). To take one concrete example: a search engine's A/B test comparing the placement of advertising banners had 3,298,086 experimental subjects recruited from all over the world (Sahni and Zhang 2019). A physical RCT, in contrast, would require experimenters to recruit subjects and move them to a physical location to administer the experiment. The scale achievable with an A/B test is hard to imagine for a physical RCT. A/B test sample sizes are limited only by the number of visitors to websites. As long as the website has a large enough number of visitors, samples can be expanded with virtually no cost (Deng et al. 2013).

Due to the power granted by such large sample sizes, A/B tests can be used to detect differences between even the most similar products (Azevedo et al. 2019). This novel ability to reliably measure even minuscule product changes opens up unprecedented opportunities for incremental improvements. Douglas Bowman's words in the epigraph to this paper can be taken as an example. In the absence of a high-powered randomized controlled trial, a 0.01% increase in the number of clicks generated by the particular shade of blue of a button is extremely challenging to detect, even with a very large number of observations. A small amount of noise in the data can hide such an effect from observers, or create the illusion of such an effect where none exists. A/B testing, in contrast, measures these minuscule differences with high precision. As a result, many seemingly unimportant elements of a website, such as the different shades of colors or the number of pixels in a border wall, can be optimized with confidence.

By facilitating the optimization of fine details, A/B testing can exacerbate the existing bias towards incremental improvements in established organizations. As discussed above, managers are incentivized to avoid risk, cut slack, and maximize short term performance

(Manso 2011; Ederer and Manso 2013). Using A/B testing, managers generate many small, yet certain, improvements. Such improvements require no risk, as they are all vetted via A/B testing before adoption. Moreover, the fact that most of these improvements will be incremental is less of a problem with established firms; even a tiny improvement might result in generating millions of dollars, justifying the cost. A 0.01% increase in the ad revenue of Google, for example, equates to tens of millions of dollars–if not hundreds of millions.

Indeed, a focus on incremental ideas likely to result in marginal improvements is common in the world of A/B testing (Deng et al. 2013). Established organizations use their massive amount of website traffic to run large experiments designed to detect minuscule improvements (Azevedo et al. 2019; Sahni and Zhang 2019). Perhaps unsurprisingly, an overwhelming number of experiments fail to demonstrate a statistically significant difference from the status quo, positive or negative (S. H. Thomke 2020b; Kohavi, Tang, and Xu 2020).

# 3 Data and Empirics

The ideal empirical setting for testing the theoretical arguments has four distinct features. First, the use of randomized experiments as a selection method should be reasonably represented in the sample. Any collection of firms without a significant representation of experimentation would not be a suitable setting for this study. Second, there should be variance in the adoption of experimentation in the sample. If the firms in the sample have used experimentation since their inceptions, then comparison both within and across subjects becomes challenging. Third, experiments should be used for ideas and products that are important to organizations. In many industries, experiments are solely used for marketing campaigns and are orthogonal to the core products and innovations. Fourth, the empirical setting should permit observation of the use of experimentation.

## 3.1 Empirical Setting: US Newspaper Industry

Based on the conditions above, newspaper websites represented an ideal setting for the study the influence of the adoption of experimentation. First, the use of experimentation is widespread in the newspaper industry, which has a higher level of A/B testing penetration than most other industries (Vo 2016). The popularity of A/B testing is also demonstrated in figure 7. Second, because the ability to run experiments is relatively new for newspaper companies, their experience serves as an example of how adopting digital tools can transform an established industry. Third, newspapers use experimentation mainly to develop their websites. In the last decade, websites became important revenue-generating products for newspaper firms (Aral and Dhillon 2020). While in many other industries websites are solely advertising tools, newspapers view websites as a core product, and they care about website performance. *The New York Times*, for example, has invested heavily in its web services to offset the drop in printed advertising revenue (Doeland 2019). Finally, the characteristics of the industry allow for easy observation of the use of A/B testing, as I detail below in section 3.4. Since U.S. newspaper websites align well with all four key features needed for this study, they were chosen as the empirical setting.

I used the Alliance for Audited Media database to construct the list of active US daily newspapers. The Alliance for Audited Media is a North American non-profit industry organization founded in 1914 by the Association of National Advertisers to help ensure media transparency and build trust among advertisers and media companies. Originally known as the Audit Bureau of Circulations, the Alliance for Audited Media is a source of verified media information and technology platform certifications, providing standards, audit services, and data for the advertising and publishing industries. At the time of this study, the database of the Alliance for Audited Media indicated that 536 of the active daily US newspapers they audit had website versions.

## 3.2 Sample Construction: Internet Archive's Wayback Machine

In order to analyze the relationship between the use of A/B testing and those newspapers' websites, I constructed a historical archive of newspaper websites using the Internet Archive's Wayback Machine. The Internet Archive is a non-profit organization that works with thousands of other organizations and individuals around to world to save copies of web pages and build an archive of the entire internet. Their current archive includes more than 330 billion web pages. These archived pages have been made publicly available through the Internet Archive's Wayback Machine. Visitors to the Wayback Machine can type in a URL, select a date range, and then access an archived version of the web page (The Internet Archive 2018).

Using the Wayback Machine's web archive API, I scraped different versions of newspaper websites over time and constructed a panel data set ranging from 1 January 2014 to 1 January 2019. Out of the 536 US newspaper web pages, 297 had at least one snapshot on the Wayback Machine for each quarter between 1 January 2014 and 1 January 2019. To construct a balanced panel, I only included in the sample those websites archived at least quarterly by the Wayback Machine. In total, 297 websites and 5,940 unique website snapshots comprise the data.

To demonstrate that the sampling did not constitute an obvious bias, I mapped the geographical distribution of newspapers by state. The distribution of the 297 newspapers correlates roughly with state population. As shown in figure 2, a larger share of these newspapers are attributed to the most populous states (including states such as California, Florida, Ohio, and Pennsylvania), and no newspapers in the sample are attributed to the comparatively low-population states of Wyoming, Vermont, and South Dakota.

—INSERT FIGURE 2 ABOUT HERE—

## 3.3 Dependent Variables: Website Similarity Using HTML and CSS

To clarify what is meant by *incremental* versus *radical* change, I have included the website transformations in figure 3 and figure 4 as illustrations. The two versions of *The Washington Times* website (figure 3) demonstrate a minor transformation, whereas the two versions of *The Hill* (figure 4) reflect a complete transformation of the website. It is important to note here that a newspaper can undergo an even more radical transformation when, for example, a mobile version is launched or the type of news reported is altered. However, the use of experimentation and A/B testing in such contexts cannot be measured. Hence, such transformations are excluded from this article.

—INSERT FIGURE 3 ABOUT HERE—

—INSERT FIGURE 4 ABOUT HERE—

Capturing and quantifying change in websites is a challenging task, and this study required a method that could reliably measure change and distinguish the incremental from the radical. A large proportion of the articles within innovation literature rely on conventional data sources and variables such as patents and R&D spending, which have minimal or no value in the context of websites. Consequently, a novel method is required to capture the degree of transformation of websites.

To solve the problem of measurement, I use a novel method that takes advantage of the underlying HyperText Markup Language (HTML) and Cascading Style Sheets (CSS) of websites. This method enabled me to use the underlying code to calculate the degree of transformation in websites over time. Although this method is novel within innovation literature, analyzing website changes based on HTML and CSS is common among web intelligence firms such as Trackly, VisualPing, Watchete, ChangeTower, and Distill.io.

—INSERT FIGURE 5 ABOUT HERE—

The website similarity measure was constructed following a paper by Gowda and Mattmann (2016) and consists of two parts that are aggregated to a single similarity score. First, I calculated structural similarity by using HTML, the standard markup language designed to display web pages. HTML is meant to offer the rendering engines clues to structure the content on a web browser. For example, when a website must display a video, it wraps the content around a <video/> tag. HTML is a document object model (DOM): a labeled and ordered tree structure in which every node represents an object. A representation of simple HTML code as a DOM can be seen in figure 6. Simply, the HTML code of a website defines the visual structure and specifies what is displayed where.

—INSERT FIGURE 6 ABOUT HERE—

To measure the structural similarity between two HTML files, I used the tree edit distance. The tree edit distance between ordered labeled trees is the minimal-cost sequence of node edit operations that transforms one tree into another (Bille 2005). Three different edits constitute the total tree edit cost. Where cost is denoted by $\gamma_{total} = \gamma_{insert} + \gamma_{remove} + \gamma_{update}$. These costs are defined per the fundamental paper by Zhang and Shasha (1989) and taken as equal.

The larger the tree edit distance between two versions, the smaller the structural similarity of two webpages (Gowda and Mattmann 2016). Let $T_i$ denote the tree structure of website $i$, and $treedistance(T_i, T_j)$ denote a function that calculates the tree distance between website $i$ and website $j$. Where $\gamma_{max}(T_i, T_j)$ is defined as the maximum possible $\gamma_{total}$ for the tree structures of website $i$ and $j$, structural similarity is defined as:

$$Structural\_Similarity = 1 - \frac{treedistance(T_i, T_j)}{\gamma_{max}(T_i, T_j)}$$

Second, I calculated the style similarity of two web pages by using CSS, a stylesheet language used to describe the presentation of a document written in HTML. CSS classes are used together with HTML tags to describe the style of inputs.

I treated the CSS style sheets as a set of features corresponding to each web page and used that as the set comparison (Gowda and Mattmann 2016). Then, using the sets of CSS classes, I compared the styles of two web pages by using Jaccard similarity. The Jaccard similarity coefficient of styles is computed by determining the fraction of styles overlapping both web pages. Let $D_i$ and $D_j$ be the sets of CSS classes from webpages $i$ and $j$. Then, their stylistic similarity is:

$$Stylistic\_Similarity = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$$

After calculating the structural and stylistic similarity, I followed Gowda and Mattmann (2016) and calculated the composite similarity score, which is defined as:

$$Similarity = k * Structural\_Similarity + (1 - k) * Style\_Similarity$$

Gowda and Mattmann define $k = 0.5$ but note that different values might be more appropriate for different types of websites. Following their application, I defined $k = 0.5$ to calculate the composite similarity score. The resulting website similarity score is a continuous variable ranging from 0 to 1. If the similarity between two versions of a website is close to 1, it indicates minimal change in the website. A similarity score close to 0 means that the two version are substantively different.

It is important to note that the similarity score has been developed to capture visual design similarity (structural and stylistic) and excludes anything related to website content, such as actual text and images. This decision was consciously and deliberately taken for the empirical setting of newspaper websites. First, the content on daily newspaper websites change every day. Any method that captures the transformations in content will detect a large amount of change due the nature of the product, even when there is no innovation happening on the website. Hence, it is crucial to disregard the content change to prevent this issue. Moreover, many interviews with the practitioners revealed that the overwhelming

number of experiments focus on the visual design elements and explicitly exclude website content (Vo 2016). Firms in other industries also commonly exclude content from experimentation; Netflix is one example (Gomez-Uribe and Hunt 2015).

The similarity score compares different versions of newspapers, taken at quarterly (three-month) intervals, over time. For example, the New York Times web page on 1 January 2015 was compared to the New York Times web page on 1 April 2015. This process was repeated at one-quarter intervals through the last observation in the data set. In other words, every row in the data set represents a comparison of two versions, one quarter apart, of a newspaper website. In total there are twenty quarters between 1 January 2014 and 1 April 2019, which yielded nineteen unique quarterly comparisons for each of the 297 websites in the data.

For robustness purposes, the same analysis was implemented at six-month and one-year intervals. The results were robust to the different comparison window specifications. For presentation purposes, these additional analyses are provided in the appendix.

Using the website similarity score, I constructed two dependent variables. The first one is $Radical\_Change$, which is a binary variable that indicates a similarity score below 0.10. This variable is used for detecting the number of radical transformations in the visual design of websites before and after the adoption of A/B testing. The value 0.10 was adopted following the clustering success demonstrated by Gowda and Mattmann 2016. The analyses presented in the results section were also conducted using 0.05 and 0.15 specifications for $Radical\_Change$. The results were robust to the different specifications of $Radical\_Change$. The second dependent variable is average similarity, denoted as $Similarity$, and it consists of the similarity scores of websites.

The website examples provided in figures 4 and 3 can be used to demonstrate how the similarity score functions. Figure 4 shows two extremely similar versions of the same website. The algorithm was able to capture that, calculating a similarity score of 0.99 out of 1.00 for those two pages. Figure 3 shows two pages with a similarity score of 0.01. These websites represent a radical change; with a similarity score below 0.10, they are almost completely

different from each other.

## 3.4 Independent Variable: Use of A/B Testing

It is common for firms to outsource their A/B testing infrastructure and use third-party software by companies such as Google, Optimizely, Adobe, and Mixpanel (S. H. Thomke 2020b). The cheapest testing option, preferred by most firms using third-party software, is client-side testing. In client-side testing, the experiment is run by a code snippet added to the header of a website's HTML code. These code snippets carry identifying information about the A/B testing software providers. Because the HTML is public and displays these code snippets, anyone can detect whether a website is using this A/B testing software.

Firms that specialize in web profiling, competitor analysis, and lead generation use web crawler algorithms to crawl the whole internet and profile every website based on the code snippets found in their HTML. I acquired data from a website profiler company that has tracked these HTML files across the entire internet since 2011. The company's web crawlers browse the internet and detect what specific technology websites use, when they start using it, and when they stop (if they stop).

The data collection method used by these web crawlers relies on third-party software programs running on websites. However, a third-party software program is not the only option for firms that wish to conduct A/B testing; they can choose to use in-house software programs. These in-house software programs cannot be recognized by web crawlers, since they do not carry identifying information about a known, third-party A/B testing software provider. Moreover, even when firms use third-party software, they may opt for the more expensive server-side testing over the cheaper client-side testing, for various technical reasons. Both in-house software and third-party server-side testing can present serious challenges for detecting websites that use A/B testing. For example, digital technology firms may avoid third-party software, as they already have enough computer and data science labor to set up an in-house experimentation system. In addition, richer and larger firms can more easily

afford server-side testing, which would prevent web crawlers from detecting A/B testing software on their websites.

I argue that these potential obstacles do not constitute issues for the US newspaper industry and we can plausibly assume newspapers running A/B tests are doing so with third-party software and client-side testing technology. This assumption is plausible because even the richest and the most technologically competent newspapers, such as *The New York Times* and *The Wall Street Journal*, use third-party software and client-side testing. Moreover, most newspapers do not have enough computer and data scientists in-house to build their own experimentation program, making it even more plausible that they would outsource the technology.

Another potential problem for this data set is the difficulty in identifying websites that do not actually adopt the technology, but only try it for a short time. Many firms can use promotional opportunities from third-party software providers to try out A/B testing technology for a limited period of time, but they ultimately decide not to buy it. Such "try and leave" cases would introduce noise to the data and should not be taken into account for the eventual analysis. To address that point, I removed any A/B testing software use that was shorter than six months.

—INSERT FIGURE 7 ABOUT HERE—

I merged the list of newspapers with the A/B testing technology data set acquired from the web profiler company. Using this proprietary data set, I observed which newspapers adopted A/B testing, as well as when they started and stopped. In total, 218 distinct websites out of 297 used A/B testing technology sometime between 1 January 2014 and 1 January 2019. Figure 7 charts the number of websites using A/B testing in each quarter.

## 3.5 Performance Data

The performance of a website can be a crucial influence on the evolution of the website. It is reasonable to expect firms will alter their product if its performance declines or it simply performs below expectations. Therefore, a newspaper website's performance can determine whether a newspaper decides to implement changes to its website. As such, controlling for website performance is crucial to isolate the effect of A/B testing.

—INSERT FIGURE 8 ABOUT HERE—

To that end, I used Amazon's Alexa Web Information Service to collect daily performance metrics of the newspaper websites in the sample. Alexa Web Information Service tracks the one million most popular websites in the world and ranks them according to the number of unique page visits. It contains historical information on metrics, including the number of unique page visits and the global popularity rank of websites. If a website is not popular enough to be among the one million most popular websites, its performance metrics are not recorded.

Alexa Web Information Service data goes back to 1 January 2016, so the performance data set's range is two years shorter than the website similarity data set. Moreover, 12 out of 297 websites in the web similarity data set do not consistently appear in the performance data, meaning that their global ranking sometimes fell out of the top million. Those websites have been removed from the analyses that involve performance data. As a result, the performance data set consists of 285 unique websites, observed from 1 January 2016 to 1 January 2019. To match the website similarity data, daily performance data are aggregated up to quarters. Figure 8 describes the relationship between the mean number of page visits and similarity scores.

# 4 Results

—INSERT TABLE 1 ABOUT HERE—

This section provides descriptive information about the sample and presents the fixed effect models. Table 1 provides the basic descriptive information about several variables such as similarity, radical change, duration of A/B testing use, average weekly website visits, and average weekly circulation for the whole sample. Table 2 provides descriptive statistics for the websites that adopt A/B testing and those that do not. As shown in table 2, A/B testing adopters were almost identical to the websites that did not adopt A/B testing.

—INSERT TABLE 2 ABOUT HERE—

Table 3 provides the descriptive statistics about the main outcome variables for comparisons that were observed with and without the use of A/B testing.

—INSERT TABLE 3 ABOUT HERE—

Figure 9 demonstrates the distribution of similarity scores with a histogram. The light gray columns represent comparisons that use A/B testing, while the dark gray columns represent the comparisons that do not use A/B testing. The dashed line in the graph indicates the threshold for categorizing a comparison as radical change: a similarity score of 0.10. As can be seen in figure 9, a large proportion of the observations has a high similarity score. Even though a non-trivial number of observations falls below the radical change threshold, those observations still constitute a small proportion of the total sample.

—INSERT FIGURE 9 ABOUT HERE—

I used two dependent variables to demonstrate the association between the adoption of A/B testing and website similarity: $Radical\_Change$ and $Similarity$. $Radical\_Change$ is a binary variable that indicates whether the $Similarity$ is below 0.10. The $Radical\_Change$

variable is used for measuring the relationship between A/B testing and profound changes in websites. *Similarity* is used for demonstrating the impact of A/B testing on the average similarity of websites.

For each dependent variable, I will first present raw correlations in the form of plain OLS regressions, adding progressively more variables to demonstrate the full model. Then, I will repeat the same structure, controlling for lagged performance.

$$Y_{it} = \alpha_i + \gamma_t + \beta(A/B\_Testing)_{it} + \theta(Controls)_{it} + \epsilon_{it}$$

$Y_{it}$ represents the similarity measure of interest, either $Radical\_Change$ or $Similarity$. The treatment indicator $A/B\_Testing_{it}$ indicates whether a website $i$ is using A/B testing at time $t$. The design includes website fixed effects ($\alpha_i$) to control for website-specific trends and quarter fixed effects ($\alpha_t$) to control for time-specific trends. Following Koning, Hasan, and Chatterji (2019), I control for the number of unique technologies $Tech\_Stack$ each website uses in a given quarter. In addition, I control for website performance using log number of average daily website visitors in the previous quarter.

## 4.1 Radical Change

Table 4 demonstrates the $Radical\_Change$ models with various specifications. Model 1 in table 4 displays the correlation between A/B testing adoption and radical change likelihood in the form of a plain OLS regression. There is a highly statistically significant correlation between the two, but this does not take website-specific trends into account. Moreover, the standard errors are not clustered in model 1. Model 2 adds website fixed effects, quarter fixed effects, and clusters the standard errors on the website level. Even though the effect size shrinks, it is still in the negative direction and highly statistically significant. Model 3 adds the technology stack controls, which constitutes the full model except for the performance variable. Model 3 demonstrates the negative association between A/B testing adoption and

the likelihood of radical change.

Model 4 expands model 3 by adding the logged number of page views to the model to control for performance. Introduction of the performance control increases the standard errors by 75% and shrinks the effect size nearly 15%. Nevertheless, A/B testing use still has a negative and statistically significant effect on the likelihood of radical change. Note that model 4 has fewer observations compared to the other models. This difference is the result of the performance data set's range being two years shorter than the main data set, and the unavailability of performance data for twelve websites.

—INSERT TABLE 4 ABOUT HERE—

### 4.1.1 Average Similarity

Table 5 demonstrates the *Similarity* models with various specifications. Model 5 in table 5 demonstrates the correlation between the use of A/B testing and websites' quarterly similarity scores. The raw correlation suggests that A/B testing is associated with a positive increase in *Similarity*. In model 6, I add website fixed effects, quarter fixed effects, and cluster the standard errors on the website level. The relationship is still highly statistically significant and positive. Model 7 adds the technology stack control, resulting in the full model specification except for performance controls. The results suggest a positive and a significant relationship between the use of A/B testing and website similarity.

Finally, model 8 demonstrates the full model with logged page views as a performance control. Even though the standard errors increase by 50%, the effect size stays the same and highly statistically significant. Like model 4, model 8 has fewer observations due to the limitations of the performance data set.

—INSERT TABLE 5 ABOUT HERE—

## 4.2 Graphical Demonstration

In this section, I present graphical evidence for the effect of the adoption of A/B testing on website change. Figure 10 and figure 11 are intended to demonstrate the transformation in outcome variables before and after the adoption of A/B testing. These figures consist solely of websites that have used A/B testing; websites that never used A/B testing are excluded. For every observation, I calculate how many quarters separate that observation from the adoption of A/B testing on that website. Then, I replicate model 2 and model 6, but in place of the A/B testing treatment indicator, the number of quarters before or after adoption is used as the main regressor.

—INSERT FIGURE 10 ABOUT HERE—

The quarterly coefficients for *Radical_Change* are presented in figure 10 and *Similarity* is presented in figure 11. The figures demonstrate that the effects we observe on A/B testing adopters come from post-adoption periods and that the coefficients from pre-adoption periods are indistinguishable from zero.

An important thing to note in both figures is that the effect of A/B testing does not appear immediately after its adoption. For *Radical_Change*, three quarters are necessary to observe effects that are indistinguishable from zero. For *Similarity*, the required time is two quarters. Furthermore, effect sizes gradually increase with the number of quarters after the adoption of A/B testing.

Taken together, these two observations are crucial pieces of evidence against alternative explanations to the study findings. One alternative explanation may be that the adoption of A/B testing is accompanied by an intentional change toward incrementalism in a firm's innovation strategy. That is, firms may know that A/B testing is more appropriate for incrementalism, and they adopt it because they have decided to focus on incremental opportunities. If that were the case, then we would expect an immediate and stable change in the level of similarity after the adoption of A/B testing. However, we instead observe a lagged

and gradual slowdown of innovative activity.

—INSERT FIGURE 11 ABOUT HERE—

Therefore, I conclude that the main prediction of this article is supported by the analyses presented above. Adoption of experimentation as a selection method in established firms results in the decline of exploratory innovation and bolsters incremental search. This is reflected here in the context of US newspaper websites and their adoption of A/B testing. Additional event study models and placebo tests are presented in the appendix as robustness checks.

# 5 Discussion

Evaluating and selecting ideas is one of the fundamental parts of the innovation process. In recent decades, we have witnessed transformations in technology and management that have made randomized controlled trials a standard tool for many firms and entrepreneurs. Encouraged by popular practitioner frameworks such as Lean and Design Thinking, more and more firms run experiments to assess their ideas instead of relying on lay theories or observational data. The speed and affordability of conducting randomized controlled trials is unprecedented, as is the statistical power achievable with digital experiments.

By providing seemingly irrefutable evidence, randomized controlled trials can be a powerful driver of innovative breakthroughs. Even the most strongly held beliefs are vulnerable when faced with the results of a well-designed, well-executed experiment. Yet at the same time, experiments can be used to test incremental improvements. Both academics and practitioners commonly criticize experimentation due to its association with incrementalism. However, this important relationship lacks empirical study. The interaction between experimentation and incrementalism offers an opportunity to study the dynamics of idea evaluation and selection within organizations. The literature on innovation has focused on

traditional evaluation methods such as managerial judgment and domain expertise; nevertheless, novel evaluation methods are becoming increasingly popular and can profoundly influence the innovation process.

This article examines the impact of experimentation on innovation when the former is used as an evaluation method. In the context of US newspaper websites, I investigated how the adoption of A/B testing technology has influenced website development over time. I implemented a novel computation method that uses the underlying HTML and CSS codes of websites to measure how much change occurs on a website in a given period. The results indicate that the use of A/B testing slows innovation by increasing the likelihood that products will stay similar over time.

In addition to offering theoretical and empirical contributions, this article contributes to the literature on innovation and search with a computational technique that allows researchers to study websites in a novel way. An overwhelming proportion of the literature relies on patents or R&D spending to measure innovative activity, which–although shown to be reliable–cannot be used for many industries. In recent decades, the digital technology industry has been arguably the most important industry both in terms of monetary value and its impact on our lives. Nevertheless, innovation literature is not equipped with the tools needed to study digital technology. This paper is one of the first to tackle this problem. The tools presented in this article, crafted specifically to work with websites, do not solve the entire problem, as we need additional tools to analyze other digital technologies such as mobile applications. Nevertheless, this article might be a first step towards developing better tools and novel methods to measure product development.

There are clear boundary conditions and limitations for this study. First, although this article presents two distinct potential mechanisms for the results, data limitations unfortunately do not allow me to test and rule out those potential mechanisms. The current data sources do not contain information on ideas that were considered for A/B tests nor on the specific experiments conducted by the studied organizations. I can only observe the outcome

of the innovation process, a limitation which prevents me from investigating potential mechanisms that propose an interaction between idea generation and evaluation in organizations. The second limitation is the endogenous adoption of A/B testing. This can constitute a problem if the adoption of A/B testing is related to the outcome variable. Even though the qualitative interviews do not indicate a clear pattern, it is still plausible to assume that the findings are produced by some form of unobserved variable bias. The third limitation of this study is that it examines a single industry. Although this is a deliberate decision made for sound empirical analysis, it is not clear whether the same effects would occur in other industries. This uncertainty presents an opportunity for future research. Fourth, the computational method used in this article does not detect or measure changes related to the content on US newspaper websites. This is also a deliberate decision, as newspapers tend to experiment on design but not content. This tendency may not hold true for other industries, or even for digitally native news websites such as Uproxx, BuzzFeed, or Vox.

Despite its limitations, this study offers several opportunities for future research. As noted in this paper, the availability of cheap digital experiments is transforming decision-making in domains ranging from strategy to human resources. The literature would benefit greatly from a closer examination of the potential mechanisms presented in this paper. These mechanisms have the potential to shed light on larger aspects of organizational decision-making, especially considering the rapid evolution of the use of data in organizations. Another question for future research is how firms might improve the impact that experimentation has on innovation. This paper does not make the argument that experimentation is destined to harm innovation, nor that it must be a net negative for firms. On the contrary, experimentation presents many important benefits to organizations for innovation, and there might be ways to ameliorate its negative impacts. Better incentive designs or organizational cultures that merge experimentation and exploration might prove vital tools for overcoming current drawbacks to experimentation in established firms (Camuffo et al. 2019; S. H. Thomke 2020a, 2020b).

# References


Aghion, Philippe, and Matthew O Jackson. 2016. "Inducing leaders to take risky decisions: Dismissal, tenure, and term limits." *American Economic Journal: Microeconomics* 8 (3): 1–38. ISSN: 19457685. doi:10.1257/mic.20150083.

Akerlof, George A. 2020. "Sins of Omission and the Practice of Economics." *Journal of Economic Literature* 58 (2): 405–418. ISSN: 0022-0515. doi:10.1257/jel.20191573.

Angrist, Joshua David., and Jörn-Steffen. Pischke. 2009. *Mostly Harmless Econometrics : An Empiricist's Companion.* First. 373. Princeton, NJ: Princeton University Press. ISBN: 9780691120355.

Aral, Sinan, and Paramveer S. Dhillon. 2020. "Digital Paywall Design: Implications for Content Demand & Subscriptions." *Management Science* Forthcomin:1–39.

Azevedo, Eduardo M., Alex Deng, José Luis, Montiel Olea, Justin Rao, and E. Glen Weyl. 2019. "A/B Testing with Fat Tails."

Berg, Justin M. 2016. "Balancing on the Creative Highwire: Forecasting the Success of Novel Ideas in Organizations." *Administrative Science Quarterly* 61 (3): 433–468. ISSN: 19303815. doi:10.1177/0001839216642211.

Bille, Philip. 2005. "A survey on tree edit distance and related problems." *Theoretical Computer Science* 337 (1-3): 217–239. ISSN: 03043975. doi:10.1016/j.tcs.2004.12.030.

Boudreau, Kevin, Eva Guinan, Karim R. Lakhani, and Christoph Riedl. 2016. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance and Resource Allocation in Science." *Management Science* 62 (10): 2765–2783. ISSN: 0025-1909. doi:10.2139/ssrn.2478627.

Bowman, Douglas. 2009. *Goodbye, Google.*

Burgelman, Robert A. 1991. "Intraorganizational Ecology of Strategy Making and Organizational Adaptation: Theory and Field Research." *Organization Science* 2 (3): 239–262. ISSN: 1047-7039. doi:10.1287/orsc.2.3.239.

Camuffo, Arnaldo, Alessandro Cordova, Alfonso Gambardella, and Chiara Spina. 2019. "A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial." *Management Science* 66 (2): 564–586. ISSN: 15265501. doi:10.1287/mnsc.2018.3249.

Csaszar, Felipe A. 2018. "Limits to the wisdom of the crowd in idea selection." *Advances in Strategic Management* 40:275–297. ISSN: 07423322. doi:10.1108/S0742-332220180000040010.

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker. 2013. "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data." In *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining,* 123–132. ISBN: 9781450318693. doi:10.1145/2433396.2433413.

Deniz, Berk Can, and Jesper B. Sørensen. 2019. "Organizational vs . Crowd Selection : Implications for Exploration and Exploitation."

Doeland, Denis. 2019. *The New York Times leader in the digital transformation.*

Ederer, Florian, and Gustavo Manso. 2013. "Is Pay for Performance Detrimental to Innovation?" *Management Science* 59 (7): 1496–1513. ISSN: 0025-1909. doi:10.1287/mnsc.1120.1683.

Felin, Teppo, Alfonso Gambardella, Scott Stern, and Todd Zenger. 2019. "Lean startup and the business model: Experimentation revisited." *Long Range Planning,* no. June. ISSN: 18731872. doi:10.1016/j.lrp.2019.06.002.

Fleming, Lee, and Santiago Mingo. 2008. "Creativity in new product development : An evolutionary integration." Chap. 5 in *Handbook of New Product Development Management,* First, edited by Christoph H. Loch and Strlianos Kavadias, 113–133. Oxford, UK: Elsevier Ltd.

Ganz, Scott C. 2020. "Hyperopic Search: Organizations Learning About Managers Learning About Strategies." *Organization Science* Articles i (April): 1–18.

Gibbons, Robert, and John Roberts. 2012. *The handbook of organizational economics,* 1–1233. ISBN: 9781400845354. doi:10.1080/01446193.2013.872281.

Girotra, Karan, Christian Terwiesch, and Karl T. Ulrich. 2010. "Idea Generation and the Quality of the Best Idea." *Management Science* 56 (4): 591–605.

Gomez-Uribe, Carlos A, and Neil Hunt. 2015. "The netflix recommender system: Algorithms, business value, and innovation." *ACM Transactions on Management Information Systems* 6 (4). ISSN: 21586578. doi:10.1145/2843948.

Gowda, Thamme, and Chris Mattmann. 2016. "Clustering Web Pages Based on Structure and Style Similarity." In *IEEE 17th International Conference on Information Reuse and Integration.*

Howard, Edward R. 2013. "Joseph Lister: His contributions to early experimental physiology." *Notes and Records of the Royal Society* 67 (3): 191–198. ISSN: 17430178. doi:10.1098/rsnr.2013.0029.

Kaplan, Abraham. 2009. *The Conduct of Inquiry.* Fourth. London, UK: Transaction Publishers. ISBN: 0765804484.

Keum, Dongil D, and Kelly E See. 2017. "The influence of hierarchy on idea generation and selection in the innovation process." *Organization Science* 28 (4): 653–669. ISSN: 15265455. doi:10.1287/orsc.2017.1142.

Knudsen, Thorbjørn, and Daniel A Levinthal. 2007. "Two Faces of Search: Alternative Generation and Alternative Evaluation." *Organization Science* 18 (1): 39–54. ISSN: 1047-7039. doi:10.1287/orsc.1060.0216.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments.* First. 1–200. Cambridge, UK: Cambridge University Press. ISBN: 1108724264. doi:10.1017/9781108653985.

Kohavi, Ron, and Stefan Thomke. 2017. "The surprising power of online experiments: Getting the most out of A/B and other controlled tests." *Harvard Business Review,* no. September-October: 1–9. ISSN: 00178012. doi:10.1017/S0266466602185070.

Koning, Rembrand, Sharique Hasan, and Aaron Chatterji. 2019. "Experimentation and startup performance : Evidence from A / B testing."

Levinthal, Daniel A. 2007. "Bringing selection back into our evolutionary theories of innovation." Chap. 9 in *Perspectives on the Economics of Innovation,* First, edited by F. Malerba and S. Brusoni, 293–307. Cambridge, UK: Cambridge University Press. ISBN: 9780511618390. doi:10.1017/CBO9780511618390.015.

Luca, Michael, and Max H. Bazerman. 2020. *The Power of Experiments: Decision Making in a Data-Driven World.* First. Cambridge, MA: MIT Press. ISBN: 978-0262043878.

Luo, Hong, Jeffrey Macher, and J. Michael Wahlen. 2020. "Judgment Aggregation in Creative Production: Evidence from the Movie Industry." doi:10.2139/ssrn.3451303.

Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance,* no. 5: 1823–1860. ISSN: 00221082. doi:10.1111/j.1540-6261.2011.01688.x.

Mollick, Ethan, and Ramana Nanda. 2016. "Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts." *Management Science* 62 (6): 1533–1553. ISSN: 0025-1909. doi:10.1287/mnsc.2015.2207.

Nelson, Richard R, and Sidney G Winter. 1982. *An Evolutionary Theory of Economic Change.* First. Cambridge, Massachusetts: Harvard University Press. ISBN: 0674272285. doi:10.2307/2232409.

Reitzig, Markus, and Olav Sorenson. 2013. "BIASES IN THE SELECTION STAGE OF BOTTOM-UP STRATEGY FORMULATION." *Strategic Management Journal* 34:782–799. ISSN: 01432095. doi:10.1002/smj.

Sahni, Navdeep S., and Charles Zhang. 2019. "Search Advertising and Information Discovery: Are Consumers Averse to Sponsored Messages?" doi:10.2139/ssrn.3441786.

Scott, Erin L., Pian Shu, and Roman M. Lubynsky. 2020. "Entrepreneurial uncertainty and expert evaluation: An empirical analysis." *Management Science* 66 (3): 1278–1299. ISSN: 15265501. doi:10.1287/mnsc.2018.3244.

Teplitskiy, Misha, Gary Gray, Eva Guinan, Hardeep Ranu, Michael Menietti, and Karim R Lakhani. 2019. "Do Experts Listen to Other Experts? Field Experimental Evidence from Scientific Peer Review."

The Internet Archive, Wayback Machine. 2018. *Wayback Machine General Information.*

Thomke, Stefan. 2008. "Learning by experimentation : Prototyping and testing." Chap. 15 in *Handbook of New Product Development Management,* First, edited by Christoph H. Loch and Strlianos Kavadias, 401–420. Oxford, UK: Elsevier Ltd.

Thomke, Stefan H. 2020a. "Building a Culture of Experimentation." *Harvard Business Review* March.

———. 2020b. *Experimentation Works: The Surprising Power of Business Experiments.* First. Boston, Massachusetts: Harvard Business Review Press. ISBN: 9781633697119.

Vo, Claire. 2016. *What Exactly Are Companies A/B Testing?* Technical report.

Zhang, Kaizhong, and Dennis Shasha. 1989. "Simple fast algorithms for the editing distance between trees and related problems." *SIAM Journal on Computing* 18 (6): 1245–1262. ISSN: 00975397. doi:10.1137/0218082.

# 6 Appendix

## 6.1 Event Study Models

The event study models only include websites that do not use A/B testing in the beginning of the panel and adopt it before the panel concludes. As a result, the number of observations drop significantly in the event study models. The models for both *Radical Change* and *Similarity* is presented below.

—INSERT FIGURE 6 ABOUT HERE—

—INSERT FIGURE 7 ABOUT HERE—

## 6.2 Placebo Test: Facebook Plug-Ins

To make sure that their content can easily be shared on Facebook, many newspaper websites use various Facebook plug-ins that allow their users to easily share the content on Facebook. These plug-ins can be detected on webpages' similar to how A/B testing software is detected.

I use the adoption of these Facebook plug-ins as a placebo test for the main models presented in table 4 and 5. Because the use of Facebook plug-ins should have no effect on the long term transformation of websites, we should observe no effect from their adoption. Table 8 and 9 display the results of the placebo tests. Even though there is a statistically significant raw correlation, the effect size and statistical significance shrink with the addition of fixed effects and the controls.

—INSERT FIGURE 8 ABOUT HERE—

—INSERT FIGURE 9 ABOUT HERE—

# 7 Tables and Figures

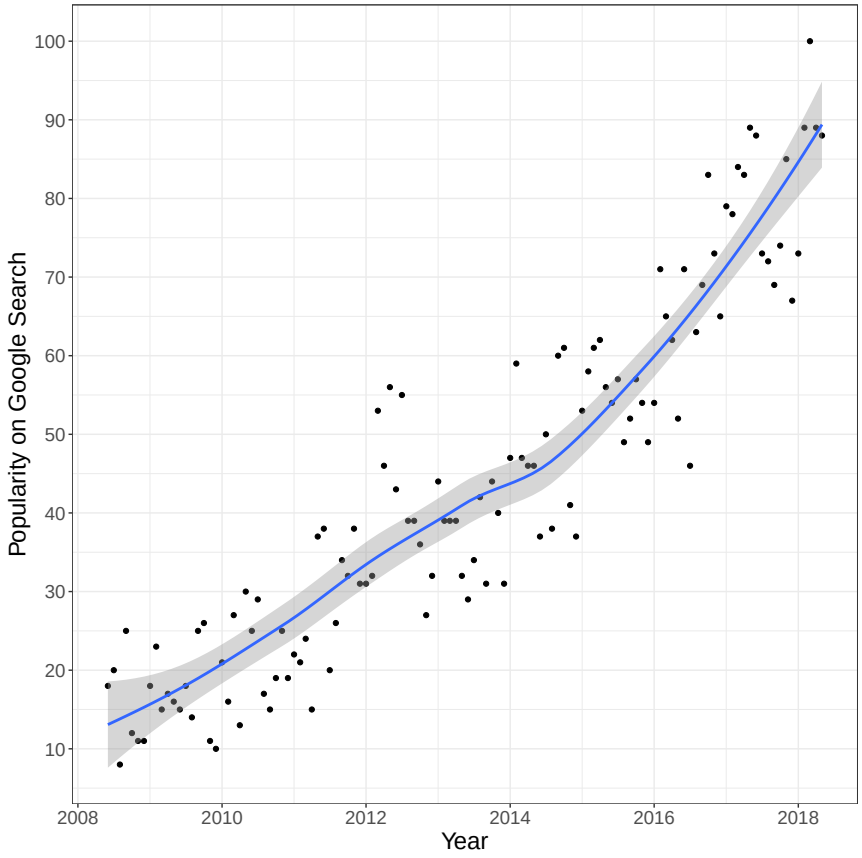


Figure 1: Worldwide popularity of the term "A/B testing" on Google Search has been increasing for the last decade

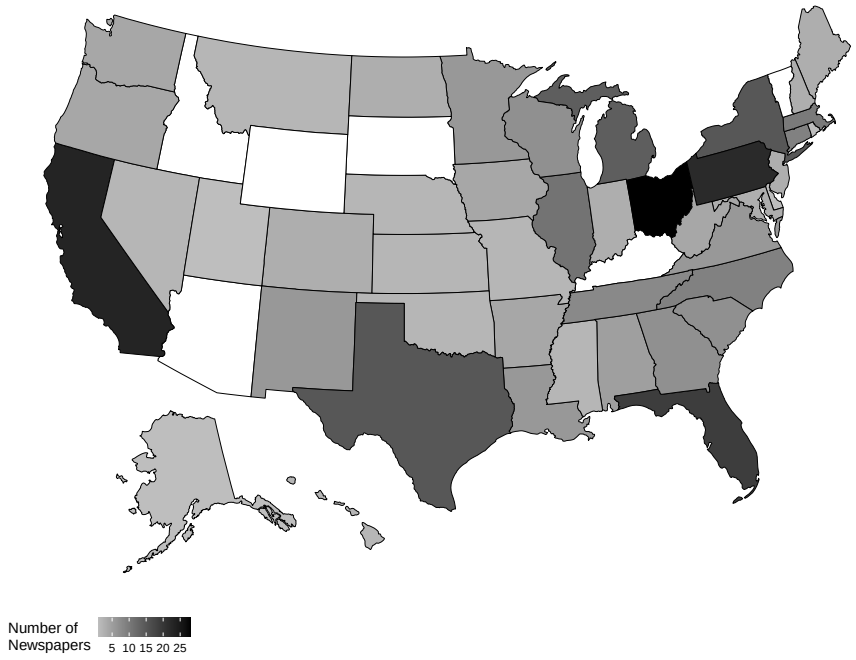


Figure 2: Distribution of the 297 newspapers in the United States

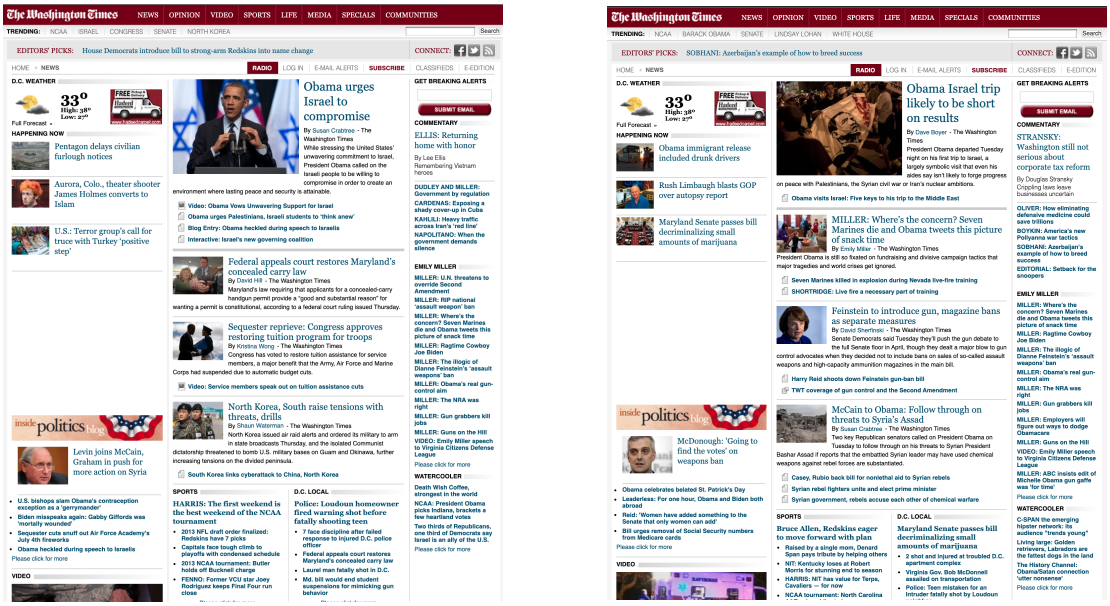

Figure 3: Two webpages with little to no change in structure and style

Figure 4: Two webpages with completely different structures and styles

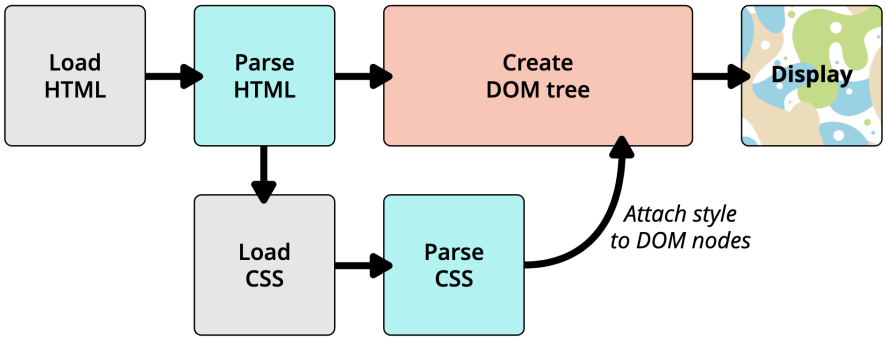


Figure 5: The process by which a web page is displayed on a local computer. CSS functions as a part of HTML and determines the styles of webpages.

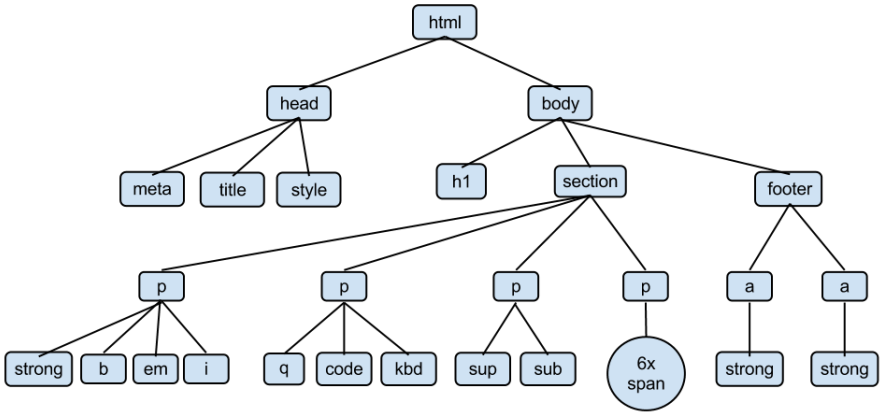


Figure 6: HTML's representation as a document object model (DOM)

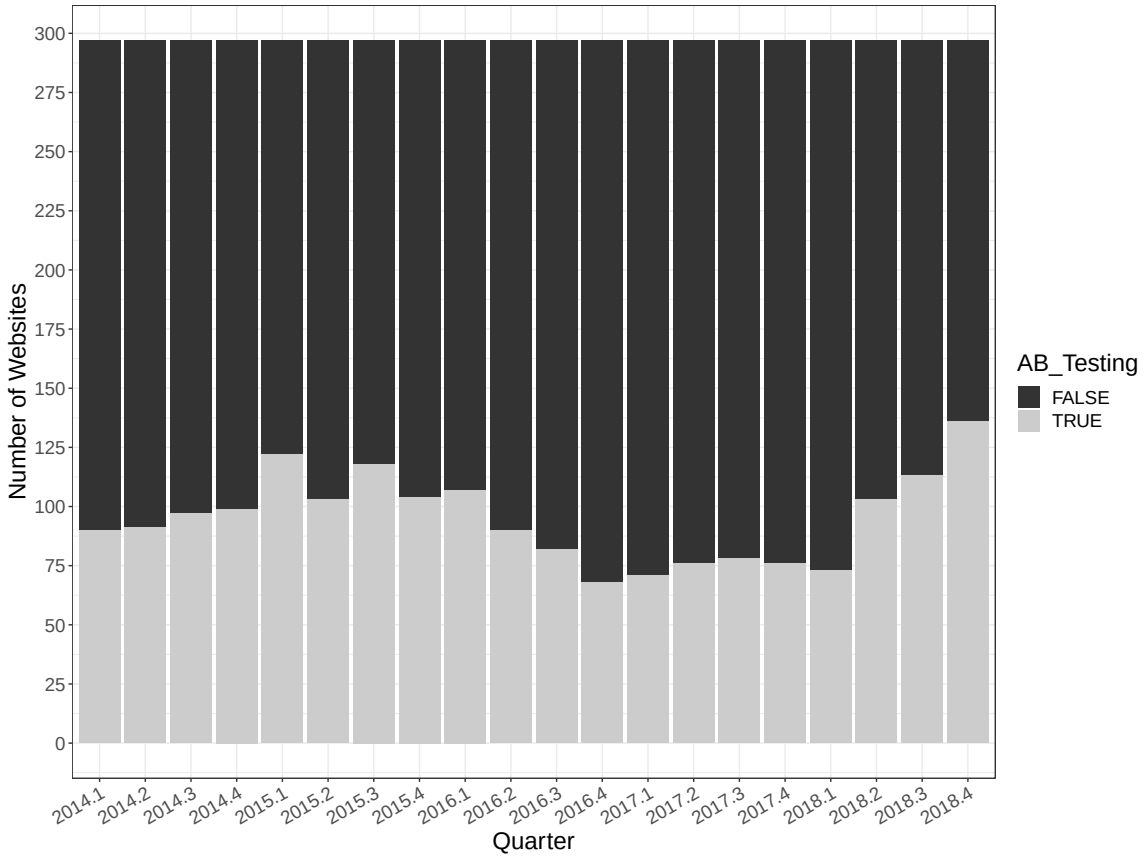


Figure 7: Number of websites using A/B testing technology in each quarter

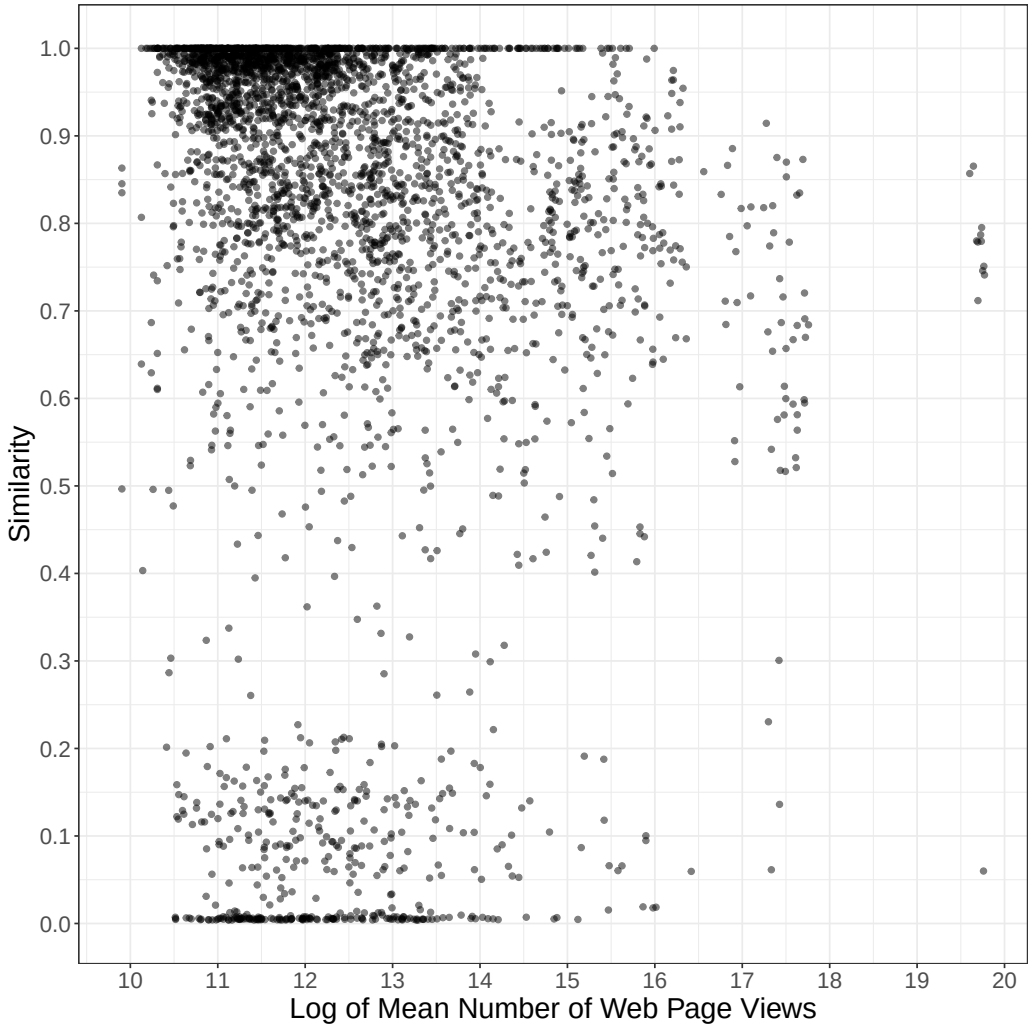


Figure 8: Scatter plot of quarterly performance measures to quarterly similarity scores

Table 1: Newspaper-level descriptive statistics

| Variable | Mean | Median | St_Dev | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| Radical_Change | 0.07 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 1.00 |
| Similarity | 0.79 | 0.88 | 0.28 | 0.00 | 0.75 | 0.98 | 1.00 |
| N_Days_AB_Testing | 791.35 | 571.00 | 552.51 | 182.00 | 328.00 | 1101.50 | 2742.00 |
| Mean_Page_View | 2619691.34 | 181220.31 | 22103576.02 | 20000.00 | 83354.78 | 539023.49 | 385870967.74 |
| Log_Mean_Page_View | 12.47 | 12.11 | 1.55 | 9.90 | 11.33 | 13.20 | 19.77 |
| Weekly_Circulation | 497122.88 | 105024.50 | 1136583.99 | 1731.00 | 28301.50 | 484587.00 | 17444147.00 |
| Log_Weekly_Circulation | 11.74 | 11.56 | 1.74 | 7.46 | 10.25 | 13.09 | 16.67 |

Table 2: Raw correlations between various newspaper-level variables and the adoption of A/B testing

| Variable | Mean | Median | St Dev | Min | P25 | P75 | Max |
|---|---|---|---|---|---|---|---|
| **Adopted A/B Testing** | | | | | | | |
| Log(Mean Website Visit) | 12.61 | 12.25 | 1.52 | 10.31 | 11.52 | 13.21 | 19.71 |
| Mean Similarity | 0.79 | 0.79 | 0.10 | 0.51 | 0.71 | 0.86 | 1.00 |
| Mean Radical Change | 0.07 | 0.03 | 0.11 | 0.00 | 0.00 | 0.05 | 0.47 |
| Log(Mean Circulation) | 11.25 | 10.73 | 1.62 | 7.54 | 10.00 | 12.53 | 15.84 |
| **Did Not Adopt A/B Testing** | | | | | | | |
| Log(Mean Website Visits) | 12.12 | 11.72 | 1.47 | 9.90 | 11.10 | 12.63 | 16.20 |
| Mean Similarity | 0.79 | 0.80 | 0.12 | 0.40 | 0.71 | 0.87 | 1.00 |
| Mean Radical Change | 0.08 | 0.05 | 0.12 | 0.00 | 0.00 | 0.08 | 0.58 |
| Log(Mean Circulation) | 11.32 | 10.73 | 1.72 | 8.77 | 10.11 | 12.72 | 14.83 |

Table 3: Descriptive information for A/B testing and non-A/B testing comparisons. These are the numbers of website comparisons, not websites themselves.

| AB_Testing | N_Comparisons | N_Radical_Change | Mean_Similarity | Median_Similarity | SD_Similarity |
|---|---|---|---|---|---|
| FALSE | 3836 | 333 | 0.77 | 0.89 | 0.30 |
| TRUE | 1807 | 52 | 0.82 | 0.87 | 0.22 |

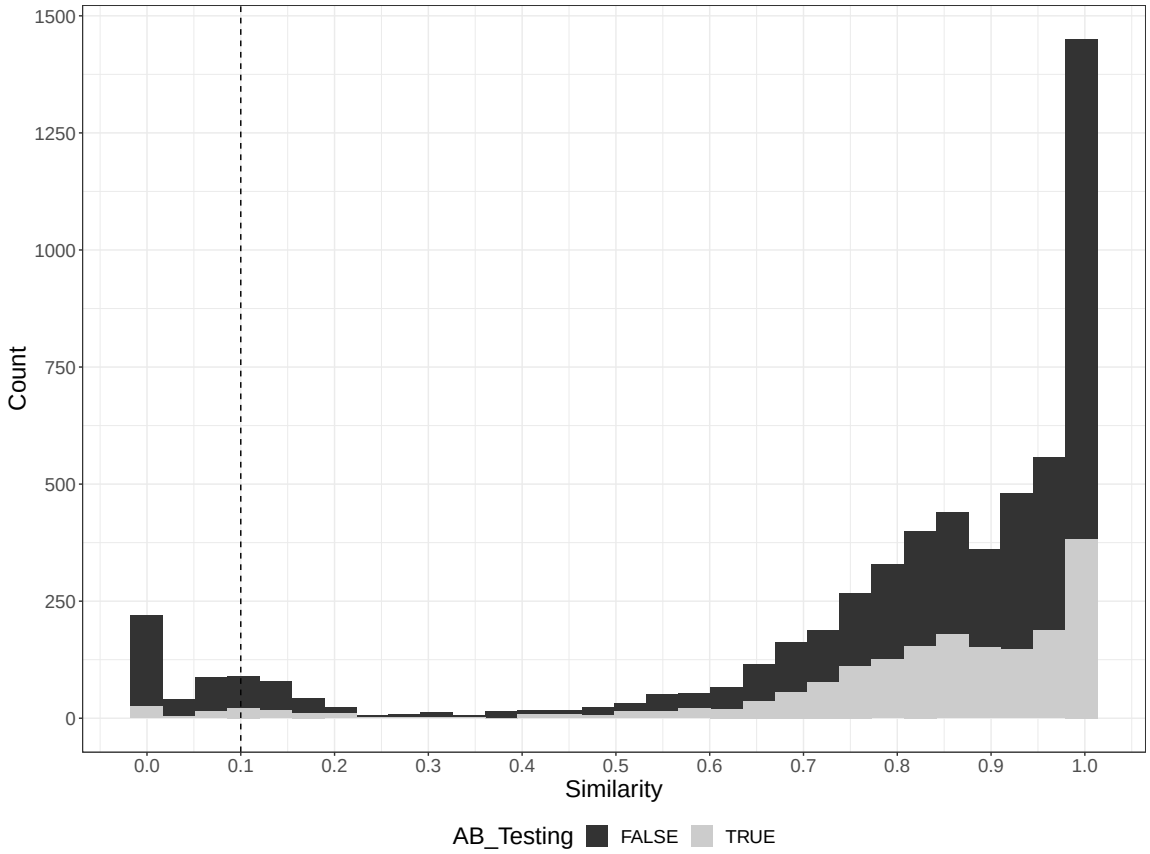


Figure 9: Distribution of Similarity. The dashed line represents the threshold for identifying radical change, 0.1

Table 4: Impact of A/B Testing Adoption on Radical Change Likelihood

| | Dependent Variable = Radical Change | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Constant | $0.087^{***}$ | | | |
| | (0.004) | | | |
| AB_Testing | $-0.058^{***}$ | $-0.048^{***}$ | $-0.045^{***}$ | $-0.037^{***}$ |
| | (0.007) | (0.009) | (0.008) | (0.013) |
| Tech_Stack | | | $-0.0004^{**}$ | $-0.001^{***}$ |
| | | | (0.0001) | (0.0002) |
| Log_Page_View | | | | $-0.008$ |
| | | | | (0.006) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 5,643 | 5,643 | 5,643 | 3,411 |
| $R^2$ | 0.012 | 0.264 | 0.265 | 0.237 |
| Adjusted $R^2$ | 0.011 | 0.139 | 0.140 | 0.164 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 5: Impact of A/B Testing Adoption on Average Similarity

| | Dependent Variable = Similarity | | | |
|---|---|---|---|---|
| | 5 | 6 | 7 | 8 |
| Constant | 0.773*** | | | |
| | (0.004) | | | |
| AB_Testing | 0.043*** | 0.067*** | 0.065*** | 0.066*** |
| | (0.008) | (0.010) | (0.010) | (0.015) |
| Tech_Stack | | | 0.0002 | 0.001*** |
| | | | (0.0002) | (0.0002) |
| Log_Page_View | | | | 0.010* |
| | | | | (0.006) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 5,643 | 5,643 | 5,643 | 3,411 |
| $R^2$ | 0.005 | 0.245 | 0.245 | 0.207 |
| Adjusted $R^2$ | 0.005 | 0.116 | 0.116 | 0.131 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

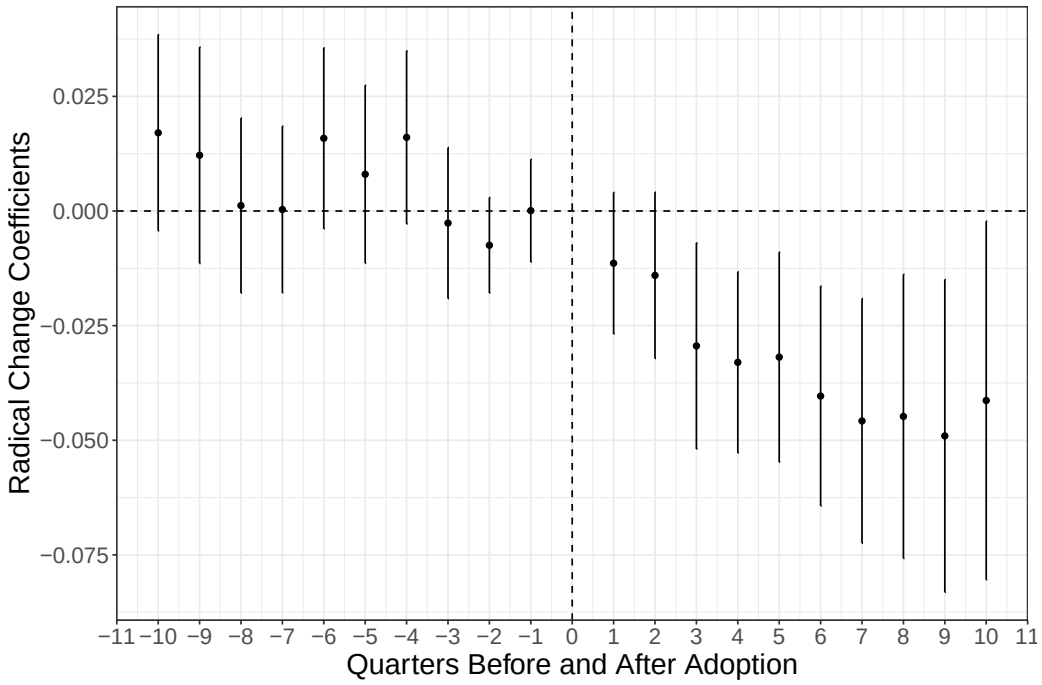


Figure 10: Radical change likelihood before and after A/B testing adoption

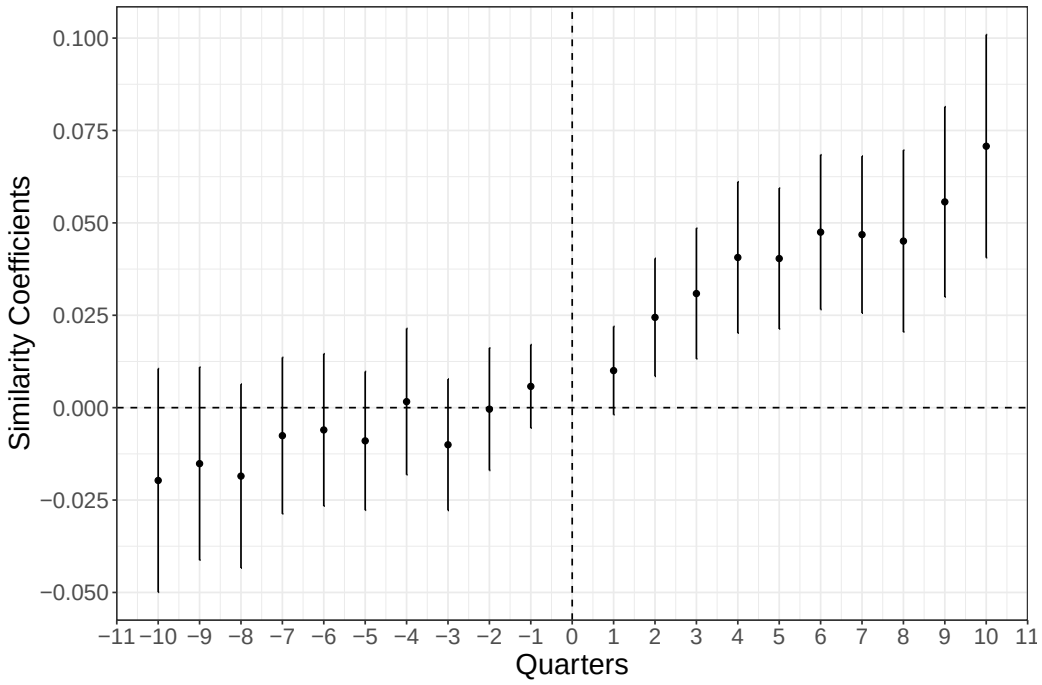


Figure 11: Average Similarity scores before and after A/B testing adoption

Table 6: Event Study Models for Radical Change

| | Dependent Variable = Radical Change | | | |
|---|---|---|---|---|
| | 9 | 10 | 11 | 12 |
| Constant | $0.125^{***}$ | | | |
| | (0.011) | | | |
| AB_Testing | $-0.094^{***}$ | $-0.063^{***}$ | $-0.063^{***}$ | $-0.157^{***}$ |
| | (0.016) | (0.021) | (0.022) | (0.056) |
| Tech_Stack | | | −0.0004 | −0.001 |
| | | | (0.0003) | (0.001) |
| Log_Page_View | | | | 0.005 |
| | | | | (0.017) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 1,159 | 1,159 | 1,159 | 468 |
| $R^2$ | 0.028 | 0.336 | 0.337 | 0.235 |
| Adjusted $R^2$ | 0.027 | 0.190 | 0.191 | 0.139 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 7: Event Study Models for Similarity

| | Dependent Variable = Similarity | | | |
|---|---|---|---|---|
| | 13 | 14 | 15 | 16 |
| Constant | 0.706*** | | | |
| | (0.011) | | | |
| AB_Testing | 0.061*** | 0.072*** | 0.071*** | 0.133** |
| | (0.016) | (0.023) | (0.023) | (0.061) |
| Tech_Stack | | | 0.0004 | 0.00000 |
| | | | (0.0003) | (0.001) |
| Log_Page_View | | | | −0.021 |
| | | | | (0.017) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 1,159 | 1,159 | 1,159 | 468 |
| $R^2$ | 0.012 | 0.288 | 0.290 | 0.199 |
| Adjusted $R^2$ | 0.011 | 0.132 | 0.134 | 0.098 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 8: Placebo Test for Radical Change

| | Dependent Variable = Radical Change | | | |
|---|---|---|---|---|
| | 17 | 18 | 19 | 20 |
| Constant | 0.029*** | | | |
| | (0.008) | | | |
| Placebo | 0.048*** | 0.013 | 0.024 | −0.060 |
| | (0.009) | (0.018) | (0.019) | (0.048) |
| Tech_Stack | | | −0.0004*** | −0.001*** |
| | | | (0.0001) | (0.0002) |
| Log_Page_View | | | | −0.008 |
| | | | | (0.006) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 5,643 | 5,643 | 5,643 | 3,411 |
| $R^2$ | 0.005 | 0.260 | 0.262 | 0.236 |
| Adjusted $R^2$ | 0.005 | 0.134 | 0.136 | 0.163 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 9: Placebo Test for Similarity

| | Dependent Variable = Similarity | | | |
|---|---|---|---|---|
| | 21 | 22 | 23 | 24 |
| Constant | $0.812^{***}$ | | | |
| | (0.009) | | | |
| Placebo | $-0.031^{***}$ | $-0.004$ | $-0.012$ | 0.034 |
| | (0.010) | (0.029) | (0.029) | (0.074) |
| Tech_Stack | | | $0.0003^{*}$ | $0.001^{***}$ |
| | | | (0.0002) | (0.0002) |
| Log_Page_View | | | | $0.010^{*}$ |
| | | | | (0.005) |
| Website Fixed Effects | No | Yes | Yes | Yes |
| Month Fixed Effects | No | Yes | Yes | Yes |
| Clustered SEs | No | Website | Website | Website |
| Observations | 5,643 | 5,643 | 5,643 | 3,411 |
| $R^2$ | 0.002 | 0.239 | 0.239 | 0.204 |
| Adjusted $R^2$ | 0.002 | 0.109 | 0.109 | 0.127 |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |