CESifo WORKING PAPERS

10457
2023

May 2023

Information Constraints and Technology Efficiency: Field Experiments Benchmarking Firms Website Performance

Anwar Adem, Richard Kneller, Cher Li

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10457*

Information Constraints and Technology Efficiency: Field Experiments Benchmarking Firms Website Performance

Abstract

This study examines the influence of information constraints on firms' efficiency in using digital technologies, focusing on business websites. Through two natural field experiments in the UK, we provide firms with benchmarked performance information on their websites. The experimental designs enable us to assess the salience of the information provided and heterogeneity linked to prior experience and catch-up potential. Our results indicate that performance gaps are not primarily driven by information constraints, as the treatment demonstrates a limited overall impact on motivating firms to improve, with a short-lived effect during the Covid-19 lockdowns. We further support these conclusions using data on website-building software and the number of page views.

JEL-Codes: O320, L250, C930, O330.

Keywords: field experiment, digital technologies, information constraints, performance management, efficiency.

Anwar Adem
Warwick Business School
University of Warwick / United Kingdom
Anwar.Adem@wbs.ac.uk

Richard Kneller
School of Economics
University of Nottingham / United Kingdom
Richard.Kneller@nottingham.ac.uk

Cher Li
Business School
University of Nottingham / United Kingdom
Cher.Li@nottingham.ac.uk

May 23, 2023

1 Introduction

The existence of substantial differences in firms' abilities to capture the full value of their investments in new technologies has been established in the literature as a robust empirical finding (Syverson, 2011). These disparities in efficiency have been found to occur both across firms in the same industry (Foster et al., 2008) and within a single firm (Bloom et al., 2015).[1] In this study, we examine the role of information constraints in explaining inefficiencies in the utilisation of digital technologies. We conduct two natural field experiments that provide UK firms with benchmarked performance information on their websites, allowing us to test, for the first time, whether information constraints can account for differences in the performance of this digital technology.

With the rise of the Internet and e-Commerce technologies, business websites have emerged as a critical component of modern firms' production functions, enabling connections with customers, shaping their online experience, and influencing sales.[2] Websites are constructed by integrating numerous software applications into stacks, alongside other intangible assets such as branding, marketing, and management. The management and performance of this asset are therefore of great importance.[3] Yet, we uncover large performance gaps between firms: for example, for two commonly used measures of website loading times, the best-performing websites (at the 90^{th} percentile) load 3 and 6 times faster than the slowest websites (at the 10^{th} percentile). The various performance metrics we use also exhibit a positive, but surprisingly low correlation. Our benchmarking treatment aims to address both the presence of these gaps and the inconsistent performance across metrics.

Past experimental evidence has demonstrated the effectiveness of peer-benchmarked feedback in motivating performance improvements for labour inputs (Gosnell et al., 2020; Cai and Wang, 2022). For example, Gosnell et al. (2020) find in a field experiment on UK airline pilots that monitoring the performance of highly skilled and trained workers leads to efficiency improvements in their flying behaviours.

[1] For instance, Foster et al. (2008) find that the top-performing firms (those at the 90^{th} percentile) produce 3 to 4 times the same quantity of output from the same inputs as those with the lowest productivity (those at the 10^{th} percentile) even for homogeneous products produced using identical technology.

[2] Data from the Office for National Statistics (ONS) for the UK suggests that, in 2019, 83% of businesses had a website, up from 70% in 2007. Data from the US Census Bureau of the Department of Commerce show that in the five years leading up to 2017, e-Commerce sales as a percentage of total retail sales in the US increased from around 5% to 13%.

[3] Anecdotal evidence suggests that performance metrics like those used in this paper correlate with the user experience. For example, Aberdeen Group has calculated that a one-second delay in website loading time can reduce sales conversion rate by 7%, page views by 11%, and customer satisfaction by 16%. In addition to this anecdotal evidence, academic studies from the literature on information systems also support this correlation (Cao et al., 2005; Weinberg, 2000; Gallino et al., 2022).

This occurs partly due to management monitoring, but further improvements result from the use of individualised targets. Cai and Wang (2022) show that similar positive effects are possible when workers comment on their managers' performance.[4] In this paper, we investigate whether similar effects can be induced by benchmarking in the performance management of digital technologies across firms, a setting where previously documented mechanisms, such as power inequalities between managers and workers, may not motivate efforts to improve.

We implement two field experiments to study the impact of providing firms with various website performance metrics benchmarked against industry peers, including loading speed and search engine optimisation (SEO) metrics. Our experimental design shares the characteristics of a natural field experiment, arising from our ability to monitor performance unobtrusively by using web-scraping tools to collect a comprehensive set of website performance metrics for a large number of firms and the absence of selection by firms into their treatment status (Al-Ubaydli and List, 2015).

The two experiments differ in several significant aspects, enabling us to compare the results and address important issues stemming from experiments of this nature. First, the trials vary in size, allowing us to consider issues of power and external validity. The first trial involves 813 UK distilling firms, while the second trial includes 6,174 retail firms. Second, we use a block randomisation design in the second experiment to examine potential differences in response to the benchmark information depending on the initial performance (hence scope to catch up) (Griffith et al., 2009) or management practices (Bloom and Van Reenen, 2007). Third, we exploit differences in the timing of our experiments to address the potential concern that the provided benchmarked information may not be seen as salient by firms, by juxtaposing two experimental periods in and out of national lockdowns induced by Covid-19.

The improvements made by untreated firms across time provide additional evidence that website performance metrics are perceived as important. During the experiment period, which coincided with the Covid-19 lockdowns in the UK, untreated firms improved all of the different performance metrics we monitored, while for the second experiment, improvements were confined to those firms that were below average in the baseline data. These improvements across time varied across performance metrics in an intriguing way. The various metrics we use differ in the degree they are apparent to the website

[4] Blader et al. (2020) demonstrate that these positive performance effects can become negative if they oppose the underlying management values of the firm. Additional examples of the effects of peer benchmarks on firm behaviours using experimental methods are rare and have largely dealt with issues of tax compliance (Slemrod, 2019).

user or not. For example, the most straightforward and easily comparable measure of loading times is when something first appears on the screen (known as the first contentful paint). Alternative loading time metrics are less apparent from simple observation and are better measured using specialist performance-monitoring software of the type that we use to collect data. In other words, the various metrics differ in their cost of information acquisition to the website owner. [5] Among untreated firms, we found that the metrics with low acquisition costs improved to a greater extent over time (compared to the baseline standard deviation) than metrics with higher costs. This improvement was evident for the average untreated firm in the first experiment and concentrated among the firms with the most significant underperformance in the baseline data in the second trial.

Across the two experiments, the treatment itself generated little motivation to improve beyond that found for untreated firms. The most robust evidence of an effect from the treatment was confined to one performance metric in the experiment coinciding with the Covid-19 lockdowns. We note that this loading-time metric has high measurement costs but is relatively easy to improve, and therefore, this effect on treated firms is distinct compared to the pattern of improvement by untreated firms described above. One month after treatment, the estimated intention-to-treat effect for this high measurement cost metric was approximately 10% of the baseline standard deviation, persisted three months after treatment, but weakened at month 6 and was no longer significant, largely due to further improvement by the control firms. Outside of this, the effects of treatment were typically very small in absolute terms and when measured relative to the baseline standard deviation, with similar well-identified zero effects from the second trial. This pattern of small treatment effects also occurred irrespective of whether the firm underperformed in the baseline data or used analytics software, and therefore were able to monitor performance themselves, or not.

To investigate if our performance metrics might overlook firms' responses to the treatment, we utilised BuiltWith, a web-scraper tool that tracks over 50,000 web-based software applications. This data allowed us to assess software changes that could enhance functionality or performance. Specifically, we examined the adoption and use of Content Delivery Network (CDN) software, which was expressly recommended in our information provided to firms for improving performance and loading speed. We found that untreated firms using CDN software in the baseline period tended to discontinue

[5] This also explains the differences in the size of the performance gaps across metrics noted already above.

its use, while treated firms showed greater retention. However, for firms not using CDN software in the baseline, no significant differences were observed between treated and untreated firms in adoption over time.

Utilising BuiltWith data, we also investigated whether firms implemented software changes to enhance performance monitoring, a costlier and demanding but potentially more effective approach. However, our findings revealed little evidence for such improvements. In fact, treated firms without baseline performance monitoring software displayed lower adoption rates than the control group, implying that they substituted the information treatment for analytics software. Additionally, when examining the number of page views to assess potential subtle changes in text, images, or videos, we found no significant impact on this outcome among treated firms.

Our study makes several key contributions. First, it provides causal evidence on the role of information constraints in the performance management of digital technologies and the potential of benchmarked feedback in addressing these inefficiencies. It extends the literature on information constraints by investigating their role in a developed country setting (the UK), providing insights into how firms with relatively better management practices might respond differently compared to those in developing countries. Second, our study focuses on a single, precisely-defined management task—website performance monitoring—across firms, offering a novel perspective on targeted interventions in the context of digital technologies. Finally, our research extends insights into the effectiveness of ICT adoption and usage, taking into account the presence of complementary management and organisational practices.

The remainder of this paper is organised as follows. The next section reviews relevant literature and Section 3 describes our experimental design and methods. Section 4 introduces the data sources and main performance metrics along with some basic summary statistics. Our empirical results are described in Section 5. Section 6 concludes.

2 Related Literature

The question of why many UK firms allow the performance management of digital technologies to deviate from best-practice gains significance from the findings of Bloom and Van Reenen (2007). They report substantial differences in the quality of managerial practice across UK firms, accounting for approximately one-third of productivity gaps. Performance monitoring exemplifies a common maintenance task firms undertake, depending on their tangible and intangible capital assets, and represents one of the numerous management processes firms engage in[6]. Our study contributes to the broad literature on management practices by dissecting these management scores to examine whether improved information, a crucial input in decision-making, can drive firm outcomes towards best practice, specifically for one well-defined task.

By investigating whether information provision can lead managers to enhance performance using an experimental design, we build on a small body of literature employing field experiments with businesses. This literature can be differentiated by interventions targeting either the firm as a whole, thus influencing all management tasks within the firm (Bloom et al., 2013; Cai and Wang, 2022; Brooks et al., 2018; Cai and Szeidl, 2018), or specific management functions, predominantly focusing on worker management (Gosnell et al., 2020). Our experiment centers on one management task across firms.

Another distinction in this literature arises from the type of intervention, including consultancy (Bloom et al., 2013), in-class training (Allen et al., 2011), and information and benchmarking (Brooks et al., 2018; Cai and Szeidl, 2018). Scur et al. (2021) emphasise that the intervention form appears to matter in determining the effects of management practice interventions. The most substantial improvements are typically observed in interventions involving management consultants while benchmarking and information yield mixed results. One potential explanation for these differences lies in the interplay between the expected private value of the information given to firms and the sum of the information acquisition and performance improvement costs[7]. Treatments involving management consultants likely impact both information acquisition costs and effort costs to improve. In our setting, the peer benchmarks we provided treated firms eliminate the cost of information acquisition but do not influence effort costs to change. This factor may also help explain the strong effects of peer benchmarks observed by Gosnell et al. (2020).

A second related literature demonstrates, using experimental approaches, that firms in developing

[6] Monitoring and maintenance of physical capital assets were included in the bundled management consulting treatment given to Indian textile firms in the intervention detailed in Bloom et al. (2013)'s and is described in the classic model of Rust (1987) as the choice between regular maintenance and regenerative maintenance of capital. More generally, the measurement of the performance of a single piece of capital has a long history in economics. See Markham (1943); Vincent (1948) for historical examples, or Bloom et al. (2015) for a more recent example.

[7] There exists a large strategy management literature that explores the role of information in decision-making. See, for example, O'Reilly III (1982); Tushman and Nadler (1978).

countries may not only be unaware of their underperformance (Bloom et al., 2013) but also fail to act on available information due to inattention (Gabaix, 2019), past mistakes when interpreting its importance (Hanna et al., 2014), or present bias, which results in them postponing investments that could improve long-term outcomes (Duflo et al., 2011). We revisit this literature in the paper's conclusion but note that data from Bloom and Van Reenen (2010) shows UK firms are, on average, relatively well managed compared to their counterparts in developing countries. This finding implies that firms' responses to information might differ between the UK and developing nations.

Our paper also contributes to the wider literature on the effects of information and communication technologies (ICT) on firm performance. It is widely recognised in this literature that the benefits of ICT to firm performance rely on the presence of complementary management and organisational practices (Bresnahan et al., 2002; Bloom et al., 2012). To further extend insights on the management-ICT relationship, researchers face two challenges: 1) accurately measuring management practices across firms; and 2) establishing causality between changes in management or ICT and firm performance[8]. Bloom et al. (2012) make significant strides in addressing the first challenge. By combining ICT data with management practice survey data, they reveal a positive effect on productivity resulting from the interaction of ICT intensity and quality management. Bloom et al. (2012) confirm this stylised fact for a large sample of US firms, while Brynjolfsson and McElheran (2016) document the spread of data-driven decision-making in the US.

To explore causality, studies typically focus on exogenous changes to ICT adoption, using policies or shocks to the ICT infrastructure for identification (Kolko, 2012; Grimes et al., 2012; Bertschek et al., 2013; Haller and Lyons, 2015; Akerman et al., 2015; Fabling and Grimes, 2016; DeStefano et al., 2018, 2020, 2022). Conducted without corresponding measures of management practices, these studies rely on evidence of heterogeneity in outcomes to suggest the importance of effective ICT usage. In this paper, we sidestep the measurement of management practices by concentrating on a single process—website performance monitoring—and employing a randomized control trial method to establish a causal test for the importance of benchmarked information in management decision making.

[8] Brynjolfsson et al. (2021) demonstrate that poor measurement of intangible investments complementary to IT technologies leads to underestimation of productivity growth in the early years of a new general purpose technology (GPT).

3 Experimental Design and Timeline

We initially compiled baseline data on performance metrics for all websites in our sample, using the data collection process described in the next section of the paper. Based on this analysis, we designed an intervention to increase the information available to managers for decision making. The intervention provided website owners with two key components: a graphical illustration of their website loading speeds and SEO results, compared to industry averages and best practice results; and a textual explanation of each metric, the evidence of their impact on sales, and potential actions for improvement[9]. Firms were benchmarked against average and best-performing (90th percentile) firms within the same industry. The intervention language was deliberately neutral in tone. To establish trust in the information quality, we referenced the data sources and explained that it was collected during a research project on website performance, conducted by university-based researchers and funded by the UK Economic and Social Research Council (ESRC)[10].

A potential concern in this type of experiment is that the benchmarked information may not be perceived as salient by firms, leading to a lack of response. To address this issue, we leveraged exogenous variations in the timing of our two experiments, which generated differences in the importance of e-commerce channels. The first experiment coincided with the Covid-19 lockdowns in the UK[11]. During this time, many firms had to rely on their websites or e-commerce channels as their primary sales method due to office and non-essential retail closures. In contrast, the second experiment occurred after the lockdowns had ended and the economy had returned to normal[12].

A timeline of the first experiment is presented in Figure 1, while the timeline for the second experiment has a similar structure and is available in Figure B1 of the Appendix. In the first experiment, firms were randomly assigned to one of the two arms at the end of the benchmark data collection period. We collected post-treatment data at months 1, 3, and 6 after sending the information to firms, denoted t, $t \in \{1,3,6\}$. In the second experiment firms are stratified based on whether they are above or below

[9] The final format for this intervention was chosen following a pilot study on 10 randomly selected firms.

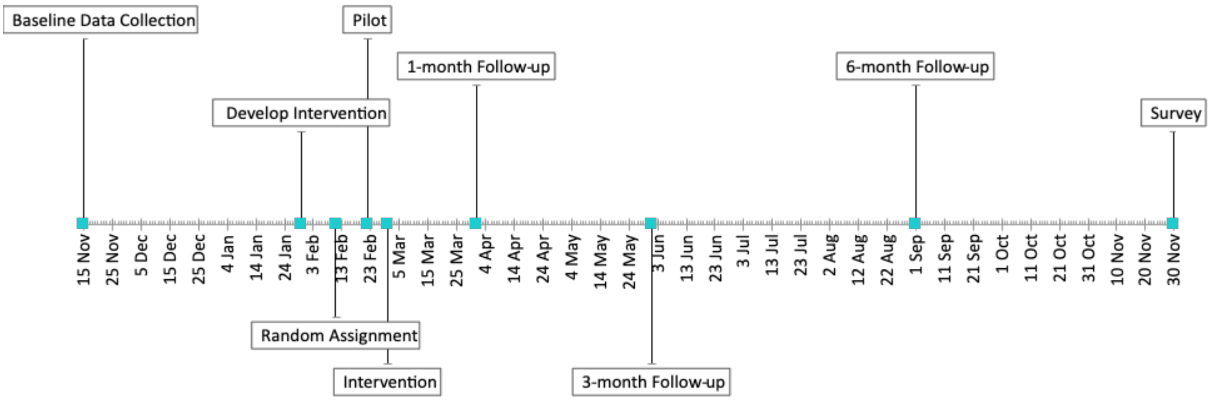
[10] The past tense was used to avoid the interpretation that firms were being actively monitored during the treatment period.

[11] The use of web scraping technologies meant that the Covid-19 period did not affect the delivery of the experiment.

[12] We also conducted a follow-up experiment with the non-treated firms from the first experiment, after the lockdowns had ended. This additional experiment involved randomly assigning firms to early and late treatment groups, using a staggered treatment design. We discuss the results of this experiment later.

average in their website performance[13] and their use of analytic software. Using these four categories of firms (above or below average and users or not of analytics software), we followed a stratified random sampling method. This approach maintains the random assignment within each category, thus the causal analysis within the bloc is as good as that overall. The timeline for the second experiment follows a parallel procedure as the first experiment with two exceptions, namely, we do not undertake a pilot study and instead of 1, 3, and 6 months, we collect data after 1, 2, and 3 months.

Figure 1: Experiment 1: Time Line



4 Data Description

4.1 Data

As a first step in the data collection process, we identify a sample of firms with websites. The first experiment uses 813 UK firms in the distilling, rectifying, and blending of the spirits industry (Standard Industrial Classification/SIC code 11010), while the second experiment focuses on 6,174 retail firms, which represents a 10% random sample of firms from this sector. We outline in full the method of identifying firms from these industries with websites in Appendix B. The second step was to measure different aspects of website performance, website traffic generated at different points in time, and software ap-

[13] We used a principal component analysis (PCA) to determine whether a firm is above or below the average Using the five dimensions of loading speed, the PCA allow us to construct a single measure summarising the five and classifying firms as above or below the average.

plications used to build the website. These all relied on web-scraping technologies and did not require contact between the research team and the firms in our sample.

The first set of performance measures focuses on website loading speed. To collect these data, we utilise an API developed by Batch Speed that searches and collects bulk data from Google's Page Speed checker, which relies on an open-source automated tool called Lighthouse for auditing web page quality[14]. We gather various loading time metrics for both desktop and mobile devices. The first measure, First Contentful Paint (FCP), records the time from the page starting to load to when any content appears on the screen. The second measure, First Meaningful Paint (FMP), indicates when the primary content becomes visible to the user. The third measure, First CPU Idle (FCPUI), represents the time it takes for a page to become minimally interactive. This measure was only available on Google Lighthouse up to three months after treatment and thus does not feature in our six-month analysis. Time to Interactive (TTI), our fourth metric, measures how quickly users can fully interact with a page. The final measure, Speed Index (SI), is a synthetic index calculated by Google Speed Checker that evaluates the average rendering speed of a web page, incorporating the previous speed variables.

By the time of the second trial, Google Lighthouse had updated some performance measures. First Meaningful Paint (FMP) and First CPU Idle (FCPUI) were discontinued, and we use Server Response Time (SRT) and Largest Contentful Paint (LCP) as similar alternatives. SRT refers to the time required for a browser to establish a connection to the server, while LCP reports the render time of the largest visible image or text block within the viewport. The other three measures remain the same as in the first trial.

The second quantifiable dimension of website performance concerns search engine optimisation (SEO), which increases a website's visibility in response to relevant user searches. Using proprietary data from a specialist company, MOZ, we collect two metrics: ranking keywords (RK), capturing the presence of frequently used keywords in web searches for that type, and domain authority (DA). The DA variable is a search engine ranking score developed by MOZ that captures website performance more broadly, based on a machine learning model that evaluates thousands of actual search results to predict search engine query outcomes, depending on keywords and loading times. DA values range from 1 to 100, with higher values corresponding to a higher likelihood of ranking well in a Google search for the

[14] Data of this type was also used by Gallino et al. (2022), Hernández et al. (2009), and Boshoff (2007).

product category. Lastly, we collect a direct measure of website traffic in the form of page views (PV) obtained from Ubersuggest, an online website monitoring, and search engine optimization platform.

Information on the software applications used to build each website was collected from https://builtwith.com/. BuiltWith is an online platform that tracks software technologies and sells this information to website software companies for lead generation, sales intelligence, and market share analysis. The set of elements used to develop an application, such as database, back-end frameworks, and front-end frameworks, is colloquially known as a "technology stack." BuiltWith captures the use of over 53,000 different web technologies, which are grouped into categories such as analytics, content management systems, content delivery networks, and web hosting providers.

4.2 Baseline Summary Statistics

Summary statistics of the main desktop loading time variables included in the first experiment are provided in Table 1, grouped into individual and overall measures of website performance. Metrics included in the second experiment cover both desktop and mobile loading times and are reported in Table B1.

We identify three stylised facts from these data. First, within the broad categories of loading time and SEO, the various metrics capture different aspects of website performance. For instance, in the first experiment, the average time for a website to generate any content (FCP) is 1.4 seconds (s.d. 0.7), while the time for the website to become interactive (TTI) averages 3.2 seconds (s.d. 2.2).

Second, there is considerable heterogeneity in website performance. Focusing on the initial (FCP) and latest (TTI) metrics of loading time from our first experiment, the fastest websites take just 0.2 seconds to generate any content (FCP), while the slowest takes 7.5 seconds. For the TTI measure, the minimum values were both 0.2, while the maximum value was 20.1. Comparing the ratios of these different website speed variables at the 10^{th} and 90^{th} percentiles shows that the performance gaps in the sample increase as we move across these two measures. For the FCP, the 90:10 ratio is close to 3:1, while, for the TTI, it is closer to 6:1. These performance gaps are even larger for the SEO performance metrics. The data for the second trial suggest that on average firms in this sector have somewhat faster websites with better SEO metrics, but the above two stylised facts remain.

Table 1: Experiment 1: Summary Statistics

	count	mean	sd	min	max	p10	p25	p50	p75	p90
First-Contentful-Paint	785	1.408	0.744	0.200	7.5	.8	.9	1.2	1.7	2.3
First-Meaningful-Paint	785	1.554	0.833	0.200	7.5	.8	1	1.3	1.9	2.5
First-CPU-Idle	785	2.490	1.485	0.200	11.2	1	1.5	2.2	3.1	4.2
Time-to-Interactive	785	3.189	2.168	0.200	20.1	1	1.7	2.8	4.1	5.8
Ranking Keywords	813	24.910	143.485	0.000	2400	0	0	1	5	24
Speed-Index	785	3.268	2.127	0.200	19.9	1.3	1.9	2.7	4.1	6.1
Domain Authority	813	17.114	14.036	1.000	69	1	6	13	25	38
Page Views	779	4368.6	46671.2	0.000	1200000	0	9	121	668	3251
Below Average Count	813	2.365	1.529	0.000	5	0	1	2	4	4
Observations	813									

A third stylised fact is the inconsistencies in performance across metrics. In Table B2, we report the pairwise correlation matrix for the observations from the first experiment. This table shows high correlations between pairs of speed measures that capture similar outcomes, such as FCP and FMP (correlation: 0.95), and First CPU Idle and Time to Interactive (correlation: 0.86), but much lower correlations outside of these. For example, the correlation between FCP and TTI is 0.57. The correlations between loading time and SEO metrics are even lower. Table B3 presents the equivalence for the second trial.

We further examine these patterns by counting the number of individual performance metrics for which the firm has worse than average performance. For the metrics used in the first experiment, the minimum value for this variable is zero, and the maximum is 5. As Table B4 in the Appendix shows, the distribution is relatively flat between these different values. Only a small subset of firms operate websites that perform consistently well on all measures (14% have a website that is above average for each metric), or consistently badly (7.5% are below average on all performance metrics). The remainder displays inconsistent performance across metrics, while the average firm shows below-average performance on between two and three different metrics. Again, these patterns are similar to observations from the second experiment, as reported in Table B5 and Table B6 in the Appendix.

To comprehend the second and third stylised facts, it is crucial to recognise the differences in the cost of information acquisition for website owners across various metrics. For instance, FCP is a straightforward and easily comparable measure that indicates when something first appears on the screen. In contrast, other metrics, such as TTI, are more accurately measured using specialist performance-monitoring software like the one we employed to collect data. These gaps imply that firms may lack comprehensive performance monitoring processes capable of identifying and rectifying under-performance across all metrics, ultimately affecting the overall performance of the website.

5 Results

5.1 *Experiment 1: Changes by Counterfactual Firms*

Table B7 shows the summary statistics for the control and treated groups from the first experiment separately.[15] The last column of the table presents the difference in the mean values between the treated and control groups and the corresponding t-test statistics of a balance test. In all cases, we find no statistically significant differences between the treatment and control firms in the pre-treatment period.

To evaluate the effect of our treatment, we estimate the following regression.

$$\Delta WP_i = \beta_0 + \beta_1 treated_i + \gamma_r + \varepsilon_i. \tag{1}$$

where ΔWP_i is the change in the web performance metric compared to the baseline time period for firm i; $treated$ is an indicator of being in the group receiving the information treatment; β_1 is the estimated coefficient for the intent-to-treat effort; γ_r denotes the regional fixed effect; and ε_i is the error term. We measure changes over time rather than a simpler t-test on the level of each performance metric in order to show how the websites of un-treated firms changed over time, which is captured by the constant term β_0.

The results from the first experiment, measuring changes in various performance metrics between the baseline and one month after treatment, are presented in Table 2. Initially focusing on cross-time changes in website performance for untreated firms using the constant term in the regression, we observe that loading speeds improved (loading times decreased) compared to their pre-treatment values. Proportionally, these improvements were most pronounced for the time it took for the first content to appear on the website (FCP) and when the browser first displayed content that users found useful (FMP). For

[15] Table B8 shows the summary statistics for the second experiment by treatment status, and Table B9 presents the same statistics for each block our stratification which we discuss below.

FCP, the results comparing one month after treatment to the baseline show a reduction in loading times of 0.25s, which represents 33% of the baseline standard deviation for this variable (the standard error is 0.05 standard deviations). For FMP, the comparable figures represent an improvement in loading times of 0.32s, accounting for 38% of the baseline standard deviation (the standard error is 0.05 standard deviations). Improvements in loading times for FCPUI and TTI are similar when measured in seconds, at 0.26 and 0.38 seconds respectively, but smaller when compared to the baseline standard deviation, at 26 and 18% respectively (the standard errors are 0.06 and 0.04 standard deviations).

These differences in the magnitude of improvement are intriguing, as they align with differences in the costs of acquiring information across website loading time measures. The most straightforward measure of loading times to evaluate is when something first appears on the screen (first contentful paint), which is easily comparable whenever a website loads. Other loading time metrics are less evident from simple observations and are better collected by performance monitoring software like the ones we use, often available if a website builder platform is utilised to construct and host the website. Improvements in performance metrics were greatest for those metrics with the lowest costs of acquiring information.

In addition to differences in the cost of acquiring information on performance, website metrics also vary in the effort required to generate improvements. Typically, these costs are lower for loading time measures and higher for search engine optimisation metrics. To enhance SEO scores, firms may need to hire specialists from a vast industry of web programmers, designers, and digital advertisers. In line with the greater costs of effort to improve, only minor enhancements in ranking keywords occurred for untreated firms over the first month of this experiment. The constant term for this variable is positive but less than 1% of the baseline standard deviation, and this change is not statistically significant from zero at conventional levels.

5.2 *Experiment 1: Immediate treatment effects*

The treatment variable in the regression suggests a weak response to the benchmarking information sent to firms. It is important to recall that this benchmarked information was disseminated during a period when e-commerce channels held greater relative importance for businesses in this sector. For the FCP measure, the magnitude is unexpectedly positive but very small (0.007s), and the standard error (0.04s) is also small, indicating a well-identified zero effect among treated firms. A similar outcome is observed

Table 2: Experiment 1: Results for Change by One-month Post-treatment

	(1) ΔFCP	(2) ΔFMP	(3) $\Delta FCPUI$	(4) ΔTTI	(5) ΔRKW	(6) ΔSI	(7) ΔDA
Treated	0.00748	0.0309	-0.0607	-0.183*	1.455	-0.278**	0.0270
	(0.039)	(0.043)	(0.075)	(0.108)	(3.156)	(0.110)	(0.131)
Cons	-0.245***	-0.315***	-0.256***	-0.382***	1.033	-0.295***	0.0973
	(0.035)	(0.040)	(0.059)	(0.091)	(1.473)	(0.084)	(0.102)
R^2	0.00875	0.0105	0.00352	0.00713	0.000989	0.0163	0.00382
N	769	769	769	769	812	769	813

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. FCP, FMP, FCPUI and TTI capture different aspect of page loading speed, denoting First-Contentful-Paint, First-Meaningful-Paint, First-CPU-Idle, and Time-to-Interactive, respectively. RKW denoting ranking keywords, captures search engine optimisation; SI stands for Speed Index and it is measured in seconds; and DA captures search engine optimisation, denoting domain authority.
*significance at the 10% level, **significance at the 5% level, ***significance at the 1% level.

for the FMP variable, where there is also a minor decrease in loading times. For the TTI and FCPUI measures, we found more substantial treatment effects, both in absolute reductions in loading time and when measured relative to the pre-treatment standard deviation. After one month post-treatment, the loading times of treated websites improved on average by 0.18s on the TTI measure and by 0.06s for the FCPUI measure. These are 0.08 and 0.04 times the standard deviation in the baseline data, while the standard errors are 0.05 standard deviations for TTI and FCPUI.[16]

In the last two columns of Table 2, we examine the overall measures of website performance. For the Speed Index variable, a similar pattern to the individual metrics above emerges. For both treated and untreated firms, there is evidence of improvement over time. Among treated firms, the SI measure improved by 0.28s, which is 0.13 of a baseline standard deviation (the standard error is 0.06 standard deviations) and is statistically different from zero at conventional levels. The Speed Index variable encompasses the improvements made to the Time to Interactive metric in Table 2. For DA, the treatment effect was positive, suggesting improvement, but the size of this effect was very small.

Why were the treatment effects confined to a small number of performance metrics?[17] An explanation consistent with this pattern of changes focuses on the costs of information acquisition and the effort required for improvements. If the cost of acquiring information on competitors was already low and thus minimally affected by the treatment, and the costs of effort to change were also low, then the incentive

[16] We also explore whether the website loading times on a mobile device responded to the treatment. These indicators were not provided to firms. The results for this can be found in Appendix C.1 and suggest similar results to those for desktop loading times.

[17] We asked this question to treated firms using a survey. Of the 29 respondents to the survey, the most common reason given was that the treatment 'acted as a reminder'. We provide more detail on this survey and its results in Appendix D.

to invest in improving that metric already existed for managers, and outcomes remained unaffected by the treatment. This arguably applies to the FCP and FMP metrics. Conversely, if the costs of effort to improve performance were high, altering the costs of information acquisition about competitors would also generate no change in managers' behaviour and, consequently, performance. This description fits the DA and ranking keywords metrics. In the final case, if information acquisition costs were previously high and decreased due to treatment, and the cost of effort to change was low, then peer benchmarking had the strongest effect.[18] This description of costs aligns with the TTI and FCPUI metrics.

5.3 *Experiment 1: Persistence*

At three months post-treatment, we observed further enhancements in website performance amongst untreated firms compared to the baseline period. As depicted in Panel A of Table 3, the reduction in the FCP loading time by three months reached 0.28 seconds, equating to 0.38 standard deviations pre-treatment, and the impact on the FMP measure was even larger at 0.43 of a baseline standard deviation. The time it takes for a website to become interactive showed improvements in loading times of 0.63s, or 0.29 times the pre-treatment standard deviation for this variable. The time it took for the first CPU to become idle improved by 0.39s, which is 0.26 of the baseline standard deviation. These improvements relative to the baseline period were still noticeable at six months amongst untreated firms, as reported in Panel B of Table 3, yet no additional enhancement was observed between the third and sixth months. For example, the FCP improvement equated to 0.36 pre-treatment standard deviations, highly similar to the change observed by the third month.[19] For TTI, the common improvements were equivalent to 0.22 of the pre-treatment value of the standard deviation for these variables, respectively, smaller than the effects at the third month. This difference between the third and sixth months amongst untreated firms is noteworthy, as the UK had fully exited the Covid-19 lockdowns by the date of this data collection in September 2021. We note that the absence of further improvement between the third and sixth months more closely resembles the results found for untreated firms in the second experiment we report below.

Examining the effects of the treatment, it is apparent that its impacts remain small, aside from the

[18] A similar conclusion that information is a necessary but often not a sufficient condition to generate an improvement in outcomes echoes a comparable finding from the literature on health outcomes in low-income countries (Dupas and Miguel, 2017).

[19] As a reminder, the FCPUI variable was discontinued by Google speedchecker by the final data collection period in month six.

effect on the TTI metric. At three months post-treatment, loading time for this metric fell by 0.21s compared to the counterfactual (i.e., a 9% change in standard deviation) and by 0.19s compared to the baseline data in the sixth month. Thus, the improvement found in treated firms in the first month persists in the third and sixth months, despite the additional reduction in website loading times by untreated firms over this period.

Table 3: Experiment 1: Results for Change by Three and Six Months Post-treatment

	(1) ΔFCP	(2) ΔFMP	(3) $\Delta FCPUI$	(4) ΔTTI	(5) ΔRKW	(6) ΔSI	(7) ΔDA
Panel A: Three Months							
Treated	-0.00903	0.0117	-0.118	-0.205*	-3.119	-0.177	-0.843
	(0.042)	(0.047)	(0.083)	(0.119)	(3.094)	(0.127)	(0.803)
Cons	-0.279***	-0.360***	-0.392***	-0.628***	3.640*	-0.613***	0.459
	(0.037)	(0.042)	(0.070)	(0.102)	(2.040)	(0.096)	(0.409)
R^2	0.00849	0.00611	0.00537	0.00693	0.00449	0.0131	0.00477
N	753	753	753	753	813	753	813
Panel B: Six Months							
Treated	-0.0281	0.00516	.	-0.194	-13.24*	-0.179	0.117
	(0.046)	(0.051)	.	(0.143)	(7.094)	(0.143)	(0.239)
Cons	-0.266***	-0.355***	.	-0.486***	10.19**	-0.480***	0.949***
	(0.041)	(0.045)	.	(0.119)	(4.902)	(0.113)	(0.185)
R^2	0.00775	0.00423	.	0.00458	0.00617	0.0121	0.0127
N	762	761	.	761	813	762	813

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. FCP, FMP, FCPUI and TTI capture different aspect of page loading speed, denoting First-Contentful-Paint, First-Meaningful-Paint, First-CPU-Idle, and Time-to-Interactive, respectively. RKW denoting ranking keywords, captures search engine optimisation. SI stands for Speed Index and it is measured in seconds. DA also captures search engine optimisation, denoting domain authority.
*significance at the 10% level, **significance at the 5% level, ***significance at the 1% level.

In alignment with the high cost of effort required to improve SEO, these measures of website performance, depicted in columns 5 and 7 of Table 3, occur more slowly. As a reminder, we previously discovered that there were no statistical differences in these performance measures compared to the baseline period after one month. Three months after the treatment, evidence from the constant in the regression indicates an increase in the number of keywords for untreated firms. Among the treated firms, there is a larger decline in the number of keywords relative to the counterfactual at three and six months. By the sixth month, treated firms used 13 fewer keywords compared to the counterfactual group, who themselves had increased the number of keywords by 10. The treated companies, therefore, appear to have responded to the treatment by making fewer textual additions to their websites. Further analysis suggests that this reduction in keywords among treated firms is driven by changes in the number of keywords amongst websites with the most keywords in the baseline data. After Winsorising the data at the 99^{th} percentile for this variable, there is no significant evidence of a decline in the number of keywords amongst treated firms.

5.4 Experiment 2: Treatment Effects and Treatment Heterogeneity

The findings from the second experiment are detailed in Table 4. Panel A outlines the change in performance when the website is accessed via a desktop device, comparing baseline data to one month post-treatment. Panel B contrasts the baseline data with two months post-treatment, while Panel C compares the baseline data with the data three months post-treatment.

Significant differences can be observed in the outcomes of this experiment in comparison to that conducted during the Covid-19 lockdown periods.[20] Initially, the constant terms indicate less systematic improvement among untreated firms over time and even some instances of performance degradation. While the Server Response Time (SRT) metric decreased by 0.05s, the Time to Interactive (TTI) metric increased by 0.08s. The variations in FCP and LCP metrics were negligible. A similar pattern is observed for these same loading time metrics measured for mobile devices in Table C3 and also occurs when we consider changes between the baseline data and two and three months post-treatment.

The treatment effects across the metrics in the Table are consistently small and statistically insignificant for all metrics, even for the time to interactive variable where significant treatment effects were apparent in the first experiment. As per the results in Panel A, the TTI metric in Table 4 decreased by 0.01 seconds (0.5% of the baseline standard deviation: s.e. 1.9%). Comparably small responses to the treatment are observed at months two and three for this variable. This pattern of results also holds when we focus on loading times metrics for mobile devices, which were also relayed to treated firms in this experiment.

Within the second experiment, we employed a block randomisation design to enable us to examine treatment heterogeneity associated with over/under-performance in the baseline data and prior experience in using website performance monitoring software. Some interesting patterns emerge from this

[20] To ascertain whether this difference emanates from disparities in the timing of the two experiments or variations in the industries – and thus the types of firms included – we conducted an additional small-scale experiment with the untreated firms from the first trial. We delve into this in more detail in Appendix E. Our overarching conclusion from this exercise is that the timing of the experiment plays a critical role in shaping the results.

Table 4: Experiment 2: Results for Desktop Devices, Changes by One, Two and Three Months Post-treatment

	(1) ΔSRT	(2) ΔFCP	(3) ΔTTI	(4) ΔLCP	(5) ΔSI
			Panel A: One Month		
Treated	-0.0245	0.00801	-0.00895	0.0145	0.00335
	(0.020)	(0.007)	(0.028)	(0.028)	(0.032)
Cons	-0.0527***	0.00265	0.0758***	-0.00962	0.0406*
	(0.015)	(0.005)	(0.022)	(0.021)	(0.023)
R^2	0.000376	0.00114	0.000586	0.000790	0.000566
N	5998	5998	5989	5993	5988
			Panel B: Two Months		
Treated	-0.00946	0.0121*	-0.00826	0.0221	-0.00130
	(0.020)	(0.007)	(0.034)	(0.031)	(0.029)
Cons	-0.0763***	0.00266	0.105***	0.00240	-0.0728***
	(0.015)	(0.005)	(0.025)	(0.022)	(0.022)
R^2	0.000358	0.00212	0.00148	0.00107	0.00130
N	5855	5855	5855	5851	5854
			Panel C: Three Months		
Treated	0.00626	0.0119	0.0372	0.0332	0.0203
	(0.021)	(0.007)	(0.033)	(0.033)	(0.029)
Cons	-0.0574***	-0.00695	-0.0271	0.00448	-0.0675***
	(0.016)	(0.006)	(0.024)	(0.025)	(0.022)
R^2	0.000410	0.00219	0.00161	0.00192	0.000791
N	5924	5924	5924	5922	5922

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed. They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index.
* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

heterogeneity. The results for loading times amongst untreated firms align more closely with the initial experiment when we consider potential heterogeneity in the treatment according to whether firms were above or below the average in the baseline data (refer to Table 5). Firms below the average among untreated firms make substantial improvements across all loading time metrics, where these effects are most pronounced for the measures of loading time that occur earliest and are most apparent to the website viewer. For instance, the SRT and FCP measures improve by 0.32 and 0.10s, equating to 36.8 and 19.7% of the baseline standard deviation (standard errors 3.3 and 1.8%), respectively. TTI and LCP show improvements among under-performing untreated firms by 0.27 and 0.33s, which represent 15.1 and 18.5% of the baseline standard deviation (standard errors are 2.7 and 2.2%), respectively. For untreated firms that were above average, hence having well-performing websites in the baseline data, we observe a significant deterioration in loading time metrics. The results, therefore, suggest catch-up by initially underperforming websites, but a worsening of performance by initially well-performing websites. If there is a failure to pay attention to website performance it is therefore concentrated amongst initially better-performing firms.

For treated firms with above-average websites in the baseline data, evidence from Panel A of the table suggests a slight deceleration in their websites post-treatment compared to the counterfactual, albeit these effects are minor, ranging between 0.03 and 0.06 seconds, and do not reach significance at conventional levels. The effects of the treatment on firms with slow websites at baseline are more frequently negative one month post-treatment, but this does not persist at two and three months. Of these metrics, the most pronounced effect is observed for the TTI metrics at month two, with a value of only -0.08 seconds.

In the second randomised control trial, we further investigated whether firms employing analytics software responded differently to information benchmarking treatment (see Table 6). Amongst untreated firms, we discern that the use of analytics software in the baseline period has no bearing on the changes in performance over time. Given that these firms could utilise the analytics software they possess to replicate the information provided in the treatment, this outcome is rather unexpected and suggests firms do not systematically leverage the available information to monitor and subsequently enhance their website's performance.[21] Consistent with this finding, the effects of the treatment are typically small in scale.

[21] The information benchmark had neither a greater nor a lesser impact on firms that use analytics software. The results for mobile device usage, presented in Table C5, indicate no robust effects. We report these results in the Appendix and do not delve into these any further in the paper.

Table 5: Experiment 2: Results for Desktop Devices, Changes by One, Two and Three Months Post-treatment by Below

	(1)	(2)	(3)	(4)	(5)
	ΔSRT	ΔFCP	ΔSI	ΔTTI	ΔLCP
			Panel A: One Month		
Treated	-0.015	-0.000	0.010	-0.009	0.021
	(0.018)	(0.007)	(0.030)	(0.029)	(0.022)
BelowAvg	-0.302***	-0.090***	-0.340***	-0.319***	-0.314***
	(0.028)	(0.008)	(0.044)	(0.041)	(0.040)
Treated×BelowAvg	-0.018	0.016	-0.013	-0.005	-0.005
	(0.040)	(0.012)	(0.063)	(0.055)	(0.054)
Cons	0.099***	0.055***	0.211***	0.240***	0.154***
	(0.014)	(0.006)	(0.020)	(0.022)	(0.016)
R^2	0.040	0.027	0.020	0.023	0.022
N	5998	5998	5988	5989	5993
			Panel B: Two Months		
Treated	0.002	0.006	-0.025	0.026	-0.001
	(0.017)	(0.008)	(0.024)	(0.033)	(0.025)
BelowAvg	-0.322***	-0.098***	-0.447***	-0.273***	-0.333***
	(0.029)	(0.009)	(0.041)	(0.048)	(0.040)
Treated×BelowAvg	-0.022	0.012	0.046	-0.079	0.050
	(0.039)	(0.013)	(0.057)	(0.067)	(0.059)
Cons	0.087***	0.059***	0.153***	0.241***	0.177***
	(0.012)	(0.006)	(0.018)	(0.023)	(0.019)
R^2	0.047	0.033	0.037	0.016	0.018
N	5855	5855	5854	5855	5851
			Panel C: Three Months		
Treated	-0.016	0.004	0.002	0.060*	0.012
	(0.018)	(0.009)	(0.026)	(0.031)	(0.026)
BelowAvg	-0.338***	-0.109***	-0.462***	-0.429***	-0.347***
	(0.030)	(0.010)	(0.041)	(0.044)	(0.047)
Treated×BelowAvg	0.044	0.016	0.037	-0.051	0.045
	(0.041)	(0.014)	(0.057)	(0.064)	(0.062)
Cons	0.112***	0.056***	0.165***	0.195***	0.186***
	(0.014)	(0.007)	(0.019)	(0.021)	(0.019)
R^2	0.038	0.033	0.040	0.034	0.019
N	5924	5924	5922	5924	5922

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed. They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index.

* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

Table 6: Experiment 2: Results for Desktop Devices, Changes by One, Two and Three Months Post-treatment by Analytic

	(1) ΔSRT	(2) ΔFCP	(3) ΔSI	(4) ΔTTI	(5) ΔLCP
			Panel A: One Month		
Treated	-0.039	0.008	-0.019	-0.001	-0.036
	(0.034)	(0.014)	(0.052)	(0.038)	(0.041)
Analytic	-0.030	-0.030***	-0.013	0.016	-0.069*
	(0.029)	(0.010)	(0.044)	(0.041)	(0.039)
Treated×Analytic	0.021	0.000	0.033	-0.011	0.074
	(0.042)	(0.015)	(0.066)	(0.053)	(0.055)
Cons	-0.032	0.023***	0.050	0.065**	0.038
	(0.023)	(0.009)	(0.034)	(0.030)	(0.028)
R^2	0.001	0.004	0.001	0.001	0.001
N	5998	5998	5988	5989	5993
			Panel B: Two Months		
Treated	-0.027	0.005	-0.050	-0.022	0.008
	(0.031)	(0.012)	(0.047)	(0.046)	(0.045)
Analytic	-0.039	-0.026**	-0.073*	0.017	-0.026
	(0.028)	(0.011)	(0.043)	(0.048)	(0.040)
Treated×Analytic	0.025	0.011	0.071	0.019	0.020
	(0.040)	(0.015)	(0.060)	(0.064)	(0.060)
Cons	-0.049**	0.021**	-0.022	0.094**	0.021
	(0.020)	(0.009)	(0.033)	(0.037)	(0.030)
R^2	0.001	0.004	0.002	0.002	0.001
N	5855	5855	5854	5855	5851
			Panel C: Three Months		
Treated	-0.039	-0.004	-0.077	0.020	-0.061
	(0.031)	(0.014)	(0.047)	(0.044)	(0.054)
Analytic	-0.038	-0.031**	-0.068	-0.001	-0.086*
	(0.030)	(0.013)	(0.043)	(0.044)	(0.051)
Treated×Analytic	0.065	0.024	0.141**	0.025	0.136**
	(0.042)	(0.017)	(0.060)	(0.062)	(0.068)
Cons	-0.031	0.014	-0.020	-0.026	0.064
	(0.022)	(0.012)	(0.035)	(0.033)	(0.043)
R^2	0.001	0.004	0.002	0.002	0.003
N	5924	5924	5922	5924	5922

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed. They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index.

* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

5.5 Experiment 1: Changes in Website Software

In this section, we consider whether website improvements were implemented by their owners but were not captured by our employed performance metrics.[22] To this end, we present evidence concerning whether website owners modified the software inputs used to build their websites, which we utilise as an indicator of their efforts to enhance their website's performance. The relationship between software and website performance is complex and highly non-linear. Certain software can append functionalities to the website, such as e-commerce or language translation, potentially slowing down speed measures, whereas others, such as CDN, are known to improve speed. Hence, we concentrate on a limited set of software types that are known to impact website performance rather than the total number of software used to construct the website.

The information dispensed to treated firms contained suggestions on how to realise improvements in website performance, one of which included the integration of CDN software into their website stack. CDNs represent a key component of the Internet infrastructure, serving as servers that replicate the content of a website's main server. These CDNs are situated in various global locations, facilitating improved website loading speeds by reducing the physical distance to the server. In Panel A of Table 7, we differentiate firms based on whether they employed CDN software in the baseline period, then observe if they added or discarded this software post-treatment. According to BuiltWith data, nearly three-quarters of firms utilise at least one CDN technology in the baseline period, with the average number of CDN software used being 1.75.

The results illustrated in Columns 1, 2, and 3 of Table 7 denote a prevailing trend towards the adoption of CDNs by untreated firms that previously did not use this type of software during the baseline period. By the six-month mark, the effect manifested significantly with a 33% increase in this likelihood. Conversely, the treatment itself had a generally negative impact on the adoption of CDNs in comparison to the counterfactual of firms that were not CDN users at baseline (Panel A, regressions 1-3), yet none of these effects are statistically significant, and inconsistencies exist within this pattern throughout the data collection period. At its peak in month six, the data implies an 8 percentage point lower probability of adoption.

Concerning firms that already utilised the technology during the baseline period (Panel A, regressions

[22] Due to funding restrictions, we were unable to gather data from BuiltWith for the firms in the second experiment.

4 to 6), the constant term intimates a common trend towards discontinuing its use. About 4 percent of untreated firms had abandoned this software by month one. By month six, this figure had risen to 6 percent. In contrast, the effect of the treatment was positive, with a coefficient effect of approximately 3%. When combined with the constant term, it seems that the treatment functioned to incentivise treated firms to abandon this software less frequently compared to the counterfactual.

In Panel B of Table 7, we next examine a set of software that we did not explicitly mention to firms in the treatment. Mobile technologies (e.g., mobile-optimised, viewport meta tag) are software packages designed to optimise website functionality on mobile devices, including enhancements to loading times. These software packages work by adjusting the website's format to reduce clutter, facilitating readability on smaller screens. For these technologies, we detected a similar pattern in the adoption and retention as with the CDN software. The constant terms in the regressions indicate an increased adoption of this mobile-optimising software among those not employing the software at baseline and some discontinuation among those using the technology during the baseline period. The treatment exhibited similar effects. We observe substantial, but poorly identified, adoption among treated firms not previously utilising the technology (Panel B regressions 1-3), and positive effects on retention (Panel B regressions 4-6). The positive effects in columns 4 to 6 dissipate more quickly than those for the CDN in Panel A, such that only the effects for month one are statistically significant. Overall, it appears that the treatment had minimal influence in encouraging firms to explore technologies that would enhance website performance beyond those explicitly mentioned in the treatment.

We can further examine software types that influence the Search Engine Optimisation (SEO) aspects of website performance. Performance monitoring software, such as A/B testing, Google Universal Analytics, and Snowplow, empowers firms to measure, collect, analyse, and report web data for the purposes of understanding and optimising web usage.[23] These tools, in addition to monitoring web traffic, assist businesses in conducting market research and enhancing website effectiveness. Technologies in this category range from A/B testing, which allows firms to compare different website versions, to Google Analytics, which provides insights into demographic, geolocation, bounce rate, click path, hit, page view, unique visitor, session, and other information derived from user interactions with the website. The second

[23] This complete set of technologies includes: GoogleAnalytics, NewRelic, Fastly, CloudflareRocketLoader, CloudflareInsights, MicrosoftApplicationInsights, AzureEdge, ReportURI, MicrosoftAzure, Heap, AkamaimPulse, Dynatrace, accessiBe, AudioEye, Site24x7, Ruxit, and GooglePageSpeedModule.

Table 7: Experiment 1: Website Software Use

	Not Adopted at the Baseline			Adopted at the Baseline		
	Mo-1	Mo-3	Mo-6	Mo-1	Mo-3	Mo-6
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. CDN						
Treated	-0.035	0.003	-0.081	0.032***	0.033**	0.029*
	(0.056)	(0.060)	(0.065)	(0.011)	(0.014)	(0.016)
Cons	0.192***	0.247***	0.335***	-0.040***	-0.047***	-0.060***
	(0.043)	(0.046)	(0.051)	(0.012)	(0.013)	(0.015)
R^2	0.015	0.023	0.019	0.018	0.017	0.007
N	215	216	214	595	595	596
Panel B. Mobile						
Treated	0.120	0.091	0.180	0.034***	0.010	0.001
	(0.095)	(0.118)	(0.118)	(0.011)	(0.011)	(0.013)
Cons	0.139*	0.358***	0.272***	-0.044***	-0.029***	-0.036***
	(0.080)	(0.101)	(0.095)	(0.012)	(0.010)	(0.011)
R^2	0.084	0.037	0.051	0.023	0.013	0.004
N	70	71	69	740	740	741
Panel C. Performance Monitoring						
Treated	-0.0633*	-0.0769**	-0.0832**	0.0253	0.0316	0.0231
	(0.033)	(0.037)	(0.037)	(0.018)	(0.022)	(0.023)
Cons	0.165***	0.200***	0.196***	-0.0494***	-0.0672***	-0.0699***
	(0.028)	(0.031)	(0.029)	(0.019)	(0.021)	(0.022)
R^2	0.0397	0.0563	0.0558	0.0152	0.00892	0.00346
N	386	387	385	424	424	425
Panel D. SEO						
Treated	0.0153	0.00978	0.0104	0.0179	0.0690	0.0568
	(0.013)	(0.015)	(0.022)	(0.031)	(0.043)	(0.043)
Cons	0.0146	0.0246**	0.0683***	-0.0665**	-0.144***	-0.123***
	(0.009)	(0.012)	(0.016)	(0.027)	(0.041)	(0.036)
R^2	0.0149	0.0172	0.0224	0.0104	0.0254	0.0312
N	612	613	612	198	198	198

Notes: Robust standard errors in parentheses.
*significance at the 10% level, **significance at the 5% level, ***significance at the 1% level.

set of software pertains to SEO optimisation technologies.[24] The use of these technologies would enable

[24] This complete set of technologies includes: Matomo, CrazyEgg, Mixpanel, Optimizely, comScore, SessionCam, Monetate, BrightEdge, SEOJSONLDBoostbyVerge, AllinOneSEOPack, SEOmatic, TheSEOFramework, RabbitSEOforWi1, YoastSEOPremium, YoastWordPressSEOPlugin, RabbitSEOforWix , SEOJSONLDBoostbyVerge, AllinOneSEOPack, SEOmatic, TheSEOFramework, RabbitSEOforWi1, YoastSEOPremium, YoastWordPressSEOPlugin and RabbitSEOforWix.

firms to monitor and enhance their website performance, which aligns with best-practice performance monitoring processes.

Panels C and D of Table 7 present the results from analysing these two sets of technologies. Some interesting distinctions emerge between these two software groups. Both for the analytic and SEO software, we observe similar evidence from the constant term as with the CDN and mobile software: an uptick in the adoption of this software group across untreated firms not utilising these technologies at baseline (Panels C and D, regressions 1-3), but some discontinuation among pre-existing users (Panels C and D, regressions 4-6). For instance, by month three, an additional 20% of firms have begun to use analytic software, while approximately 3% more firms employ SEO optimisation software.

Among the treated firms, we observe a lower adoption of this performance monitoring software among those that did not previously utilise the technology (Panel C, regressions 1-3). The effect of the treatment, as per the regressions, was to diminish the likelihood of adopting performance monitoring software between 6.3% at month one and 8.3% at month six. It thus appears that firms may have substituted the use of analytics software with the information we provided. The information treatment incorporated free sources of performance monitoring data, which could potentially elucidate this result. There was no comparable effect for SEO software. Among those already using performance monitoring or SEO software (regressions 4 to 6), we detected no evidence of a change in their usage among treated firms. For firms using SEO or analytics software at baseline, we find untreated firms discontinued their use over time, but the treatment had no effect on this decision.

5.6 Experiment 1: Page Views

In our final analysis, we investigate whether firms adjusted aspects of their website design that are challenging to quantify but could potentially influence the number of visits received.[25] Along with search engine optimisation factors such as download speed and keyword ranking, the number of page views a website attracts can be affected by the intangible dimensions of website design, including layout, ease of navigation, colour palette, and the selection of pictures and videos. In Table 8, we report outcomes for page views for firms in the first experiment at one, three, and six-month intervals. Notably, the constant across the regression table is negatively significant, suggesting a decrease in website traffic of

[25] Due to funding limitations, it was not feasible to finance data capture for this variable in the second experiment.

approximately 500 relative to the baseline among untreated firms. However, these effects are minimal compared to the baseline standard deviation of over 46,000. The effect on page views among treated firms is positive, albeit small in magnitude, and not statistically significant at standard levels. We deduce from this that, even if firms are making other changes to optimise their website performance in dimensions that we find challenging to quantify, these changes are not significantly impacting their page views.

Table 8: Experiment 1: Change in Page Views

	One Month	Three Months	Six Months
	(1)	(2)	(3)
Treated	519.5	854.8	224.5
	(901.351)	(883.148)	(929.658)
Cons	-464.9	-860.7*	-508.4
	(548.891)	(466.700)	(469.254)
R^2	0.00467	0.00451	0.00302
N	779	776	773

Notes: Robust standard errors in parentheses and estimation are cluster at firm level. Dependent variables is number of page views. Page Views (PV) measure the traffic flow of the website.
* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

6 Discussion and Concluding Remarks

The primary objective of this paper was to investigate whether information gaps or behavioural biases in the use of existing data could account for disparities in the performance of a key digital asset widely employed by UK firms. Prior literature had identified the significance of these biases in developing country contexts, while peer benchmarking had been shown to have transformative impacts on worker productivity within developed countries.

To test this hypothesis, we conducted a natural field experiment in which selected firms received peer-benchmarked information regarding their website performance. We also examined the salience of the information, using variation in the importance of e-commerce channels throughout the two experiments we conducted, and treatment heterogeneity related to initial under-performance and experience with performance monitoring software. In addition, we explored changes in the software inputs composing a website and the number of page views a website receives in order to ascertain whether treated firms responded to the treatment in ways not well captured by the performance metrics employed.

Our findings offer only weak evidence that peer-benchmark information leads to improvements in website performance, implying that information gaps do not seem to explain performance gaps in this digital technology. The exception occurs for a website performance metric with high information acquisition costs that decline as a result of treatment, and for which improvement effort costs are low. The lack of strong responses among treated firms to benchmarked peer comparisons, in contrast to earlier studies employing similar methods such as Gosnell et al. (2020), suggests the relevance of power relationships between firms and workers in those settings.

Several potential explanations for the limited response to peer benchmarking of websites could be considered. One possibility is that firms already had efficient performance monitoring systems in place, rendering the provided information of limited value. Alternatively, they might respond only if underperforming with room for improvement. Employing a block-randomisation design in which treatment is randomised according to these characteristics, we find no evidence supporting this heterogeneity. We also discovered no indication that treated firms adopted website software enabling improved performance monitoring processes.

Another explanation could be that firms do not perceive website performance or the provided metrics as important. We use differences in the significance of e-commerce channels for generating sales over time to shed light on this. The limited effects of the treatment when e-commerce channels were the primary route for firms' sales suggest that this is not a valid explanation for our results. Moreover, the fact that counterfactual firms, none of whom were aware of their participation in the experiment, demonstrated improvements in all performance measures contradicts this perspective.

A final explanation might involve other constraints or behavioural biases preventing firms from responding as expected to the information they already possess. As outlined in the literature review section of this paper, a broader literature exists on firms' failure to use the information they have access to, employing field experiments. Common explanations found in Bloom et al. (2013), Dessein et al. (2016), and Kim (2019) include inattention or failure to notice under-performance due to time constraints faced by managers. Our results provide mixed evidence for the significance of these types of behavioural biases. While untreated firms with initially slow websites managed to close performance gaps over time, suggesting that failure to notice or act is not an issue in our setting, we also found that untreated firms with better-than-average baseline websites experienced declines in performance over time. Furthermore, we observed no differences in the information treatment effect for firms with prior experience using performance monitoring software. Consequently, future research addressing behavioural biases in information utilisation appears to be a promising avenue for exploring this topic.

Bibliography

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The Skill Complementarity of Broadband Internet. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Al-Ubaydli, O. and List, J. A. (2015). Do natural field experiments afford researchers more or less control than laboratory experiments? *American Economic Review*, 105(5):462–66.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., and Lun, J. (2011). An Interaction-based Approach to Enhancing Secondary School Instruction and Student Achievement. *Science*, 333(6045):1034–1037.

Bertschek, I., Cerquera, D., and Klein, G. J. (2013). More bits–more bucks? Measuring the Impact of Broadband Internet on Firm Performance. *Information Economics and Policy*, 25(3):190–203.

Blader, S., Gartenberg, C., and Prat, A. (2020). The contingent effect of management practices. *The Review of Economic Studies*, 87(2):721–749.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does Management Matter? Evidence from India. *The Quarterly Journal of Economics*, 128(1):1–51.

Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2015). Does working from home work? evidence from a chinese experiment. *The Quarterly Journal of Economics*, 130(1):165–218.

Bloom, N., Sadun, R., and Van Reenen, J. (2012). Americans Do IT Better: US Multinationals and the Productivity Miracle. *American Economic Review*, 102(1):167–201.

Bloom, N. and Van Reenen, J. (2007). Measuring and Explaining Management Practices across Firms and Countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Bloom, N. and Van Reenen, J. (2010). Why Do Management Practices Differ across Firms and Countries? *Journal of Economic Perspectives*, 24(1):203–24.

Boshoff, C. (2007). A psychometric assessment of es-qual: a scale to measure electronic service quality. *Journal of Electronic Commerce Research*, 8(1):101.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-level Evidence. *The Quarterly Journal of Economics*, 117(1):339–376.

Brooks, W., Donovan, K., and Johnson, T. R. (2018). Mentors or Teachers? Microenterprise Training in Kenya. *American Economic Journal: Applied Economics*, 10(4):196–221.

Brynjolfsson, E. and McElheran, K. (2016). The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, 106(5):133–139.

Brynjolfsson, E., Rock, D., and Syverson, C. (2021). The Productivity J-curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics*, 13(1):333–72.

Cai, J. and Szeidl, A. (2018). Inter-firm Relationships and Business Performance. *The Quarterly Journal of Economics*, 133(3):1229–1282.

Cai, J. and Wang, S.-Y. (2022). Improving management through worker evaluations: Evidence from auto manufacturing. *The Quarterly Journal of Economics*, 137(4):2459–2497.

Cao, M., Zhang, Q., and Seydel, J. (2005). B2c e-commerce web site quality: an empirical examination. *Industrial management & data systems*.

Dessein, W., Galeotti, A., and Santos, T. (2016). Rational inattention and organizational focus. *American Economic Review*, 106(6):1522–36.

DeStefano, T., Kneller, R., and Timmis, J. (2018). Broadband Infrastructure, ICT use and Firm Performance: Evidence for UK Firms. *Journal of Economic Behavior & Organization*, 155:110–139.

DeStefano, T., Kneller, R., and Timmis, J. (2020). Cloud Computing and Firm Growth. *CESifo Working Paper*.

DeStefano, T., Kneller, R., and Timmis, J. (2022). The (fuzzy) Digital Divide: The Effect of Universal Broadband on UK Firm Performance. *Journal of Economic Geography (forthcoming)*.

Duflo, E., Kremer, M., and Robinson, J. (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from kenya. *American economic review*, 101(6):2350–90.

Dupas, P. and Miguel, E. (2017). Impacts and determinants of health levels in low-income countries handbook of economic field experiments. 2. *North-Holland edited by Abhijit Vinayak Banerjee and Esther Duflo*, 2:3–93.

Fabling, R. and Grimes, A. (2016). Picking Up Speed: Does Ultrafast Broadband Increase Firm Productivity?

Foster, L., Haltiwanger, J., and Syverson, C. (2008). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review*, 98(1):394–425.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 261–343. Elsevier.

Gallino, S., Karacaoglu, N., and Moreno, A. (2022). Need for speed: The impact of in-process delays on customer behavior in online retail. *Operations Research*.

Gosnell, G. K., List, J. A., and Metcalfe, R. D. (2020). The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains. *Journal of Political Economy*, 128(4):1195–1233.

Griffith, R., Redding, S., and Simpson, H. (2009). Technological catch-up and geographic proximity. *Journal of Regional Science*, 49(4):689–720.

Grimes, A., Ren, C., and Stevens, P. (2012). The Need for Speed: Impacts of Internet Connectivity on Firm Productivity. *Journal of Productivity Analysis*, 37(2):187–201.

Haller, S. A. and Lyons, S. (2015). Broadband Adoption and Firm Productivity: Evidence from Irish Manufacturing Firms. *Telecommunications Policy*, 39(1):1–13.

Hanna, R., Mullainathan, S., and Schwartzstein, J. (2014). Learning through noticing: Theory and evidence from a field experiment. *The Quarterly Journal of Economics*, 129(3):1311–1353.

Hernández, B., Jiménez, J., and Martín, M. J. (2009). Key website factors in e-business strategy. *International Journal of information management*, 29(5):362–371.

Kim, H. (2019). The Value of Competitor Information: Evidence from a Field Experiment. Technical report, Working Paper Harvard Business School.

Kolko, J. (2012). Broadband and Local Growth. *Journal of Urban Economics*, 71(1):100–113.

Markham, J. W. (1943). Regional labor productivity in the textile industry. *The American Economic Review*, 33(1):110–115.

O'Reilly III, C. A. (1982). Variations in decision makers' use of information sources: The impact of quality and accessibility of information. *Academy of Management journal*, 25(4):756–771.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033.

Scur, D., Sadun, R., Van Reenen, J., Lemos, R., and Bloom, N. (2021). The World Management Survey at 18: Lessons and the Way Forward. *Oxford Review of Economic Policy*, 37(2):231–258.

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, 57(4):904–54.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature*, 49(2):326–65.

Tushman, M. L. and Nadler, D. A. (1978). Information processing as an integrating concept in organizational design. *Academy of management review*, 3(3):613–624.

Vincent, P. (1948). Variations in productivity between cotton spinning mills. *Journal of the Textile Institute Proceedings*, 39(8):P407–P414.

Weinberg, B. D. (2000). Don't keep your internet customers waiting too long at the (virtual) front door. *Journal of interactive marketing*, 14(1):30–39.

A Appendix

A.1 Business Websites and Their Software Components

Business websites are generally constructed around three key parts: the shopfront, products, and checkout. The virtual shopfront is analogous to the shop window and is used to highlight key information about the brand, offers, featured products, and events. The product pages include product information and prices. The checkout allows for the completion of purchases and arrangements for payments and shipping. All firms include shopfront and product information, but not all firms include checkout functions. This might be because they choose not to sell via e-Commerce, or they allow for purchases to be made via third-party providers, such as Facebook Marketplace, Amazon, or eBay, among many others. A vast array of additional functions can be added to the website. A recent prominent example is the use of analytics software to study how customers move through a website and/or capture other data.

There are two main routes for the development of commercial websites. The first is to use a website builder or content management system (CMS) such as WordPress, Drupal, Wix, Squarespace, Weebly, and Go-Daddy Website Builder. These have templates that can be customised in a limited way, and various plans that offer a range of functions, including checkouts, designed to suit the sales volumes and number of products. The second option is to build a website from scratch using coding tools such as HTML, CSS, and JavaScript. This method offers complete firm control and customisation. Generally, in both cases, checkout or shopping and payment functions can be added either by partial coding and integration or by adding pre-programmed add-ons. Adding shopping and payment functionality creates an e-Commerce website. Platform and software providers such as BigCommerce, GoDaddy, Shopify, 3dcart, Volusion, and BigCartel provide such functionalities.

B Data Description

In the first experiment, we focus on the population of UK firms operating in the distilling, rectifying, and blending of spirits industry (Standard Industrial Classification/SIC code 11010), while our second experiment concentrates on a random sample of firms in the retail sector (SIC code 47). To identify all companies operating under these SIC codes, we utilised the Financial Analysis Made Easy (FAME) database, which contains essential legal, organisational, and financial information on UK companies derived from records they are legally mandated to submit to the UK Companies House. From FAME, we collected company names, registration numbers, addresses, and industry classifications. At the time of our baseline data collection in November 2020, there were 2,200 active firms listed in the distilling industry and 271,250 firms in the retail sector, from which we selected a 10% random sample. In both industries, the majority of firms were independent, family-owned, and often too small to be required to file full accounts (including sales, employment, and financial data) at Companies House. After excluding firms without registration numbers, our analysis encompassed 2,079 firms in the distilling industry and 27,125 firms in the retail sector.

Subsequently, we identified firms with a website using both self-reported data and manual searches. Website information was voluntarily provided by 489 firms in the contact details section within FAME. To ascertain website addresses for firms with missing data in FAME, we performed fuzzy matching of firm names and addresses using FAME and the output of Google searches, focusing solely on firms in the distilling industry. Following extensive manual verification, we identified 1,066 distilling firms with websites and 7,677 retail firms. To prevent spillover of information between treatment and control groups in the field experiment, we included only one firm from those sharing the same global ultimate owner. After applying this exclusion criterion and removing another 10 distilling firms that had been recruited for the pilot study, our final sample comprised 813 unique distilling firms and 6,174 retail firms.[26]

In the empirical analysis, we combined data from multiple sources to measure various aspects of website performance, website traffic generated at different time points, and software applications employed to build the website. All data were collected online and did not necessitate contact between the research team and the firms in our sample.

B.1 Summary Statistics

[26] No further information on firms without websites is available from the FAME database beyond their name, address, owners, and industry. The absence of this information led us to surmise that these missing firms were mostly newly founded, pre-revenue firms or operated through craft or farmers markets and other ad hoc channels.

Table B1: Experiment 2: Descriptive Statistics (Baseline)

	count	mean	sd	min	max	p10	p25	p50	p75	p90
					Desktop					
Server response time	6174	0.665	0.876	0.02	9.73	.09	.16	.37	.79	1.56
First contentful paint	6174	1.002	0.496	0.2	6.5	.5	.7	.9	1.1	1.5
Speed index	6174	2.313	1.633	0.2	14.8	.8	1.2	1.9	2.9	4.4
Interactive	6174	2.198	1.804	0.2	21.6	.7	1	1.7	2.9	4.3
Largest contentful paint	6174	2.381	1.801	0.2	31.4	.9	1.3	1.9	2.9	4.3
Total blocking time	6166	90.01	295.22	0.0	9960	0	0	10	70	220
Size Desktop	6166	3267	4498	0.0	111781	459	1088	2026	3789	6939
					Mobile					
Server response time	6174	0.541	1.074	0.01	25.79	.03	.07	.19	.61	1.32
First contentful paint	6174	3.968	2.438	0.6	40.2	1.8	2.6	3.4	4.6	6.8
Speed index	6174	8.461	6.137	0.6	111.1	2.9	4.5	7.1	10.6	15.5
Interactive	6174	12.254	9.096	0.6	116.9	3.5	6.3	10.6	15.8	22.3
Largest contentful paint	6174	9.623	7.956	0.7	138.6	3.2	4.9	7.6	12	17.7
Total blocking time	6173	879	1781	0.0	64950	10	1	380	1020	2100
Size Mobile	6173	3035	4285	0.0	111771	397	977	1867	3542	6501
Observations	6174									

Table B2: Experiment 1: Pairwise Correlation (Baseline)

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) First-Contentful-Paint	1.000							
(2) First-Meaningful-Paint	0.948***	1.000						
	(0.000)							
(3) First-CPU-Idle	0.487***	0.505***	1.000					
	(0.000)	(0.000)						
(4) Time-to-Interactive	0.568***	0.575***	0.858***	1.000				
	(0.000)	(0.000)	(0.000)					
(5) Ranking Keywords	0.043	0.047	0.097***	0.188***	1.000			
	(0.233)	(0.187)	(0.006)	(0.000)				
(6) Speed-Index	0.628***	0.637***	0.625***	0.717***	0.051	1.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.153)			
(7) Domain Authority	0.136***	0.146***	0.112***	0.172***	0.349***	0.063*	1.000	
	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.079)		
(8) Page Views	0.015	0.011	0.076**	0.154***	0.752***	0.010	0.256***	1.000
	(0.671)	(0.759)	(0.036)	(0.000)	(0.000)	(0.776)	(0.000)	
(9) Below Average Count	0.655***	0.662***	0.602***	0.609***	0.057*	0.565***	-0.009	0.048
	(0.000)	(0.000)	(0.000)	(0.000)	(0.107)	(0.000)	(0.790)	(0.178)

Notes: * Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level. Standard errors in parentheses.

Table B3: Experiment 2: Pairwise Correlation (Baseline)

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) D_serverresponsetime	1.000													
(2) D_firstcontentfulpaint	0.194***	1.000												
	(0.000)													
(3) D_speedindex	0.555***	0.563***	1.000											
	(0.000)	(0.000)												
(4) D_interactive	0.151***	0.447***	0.614***	1.000										
	(0.000)	(0.000)	(0.000)											
(5) D_largestcontefulpaintt	0.165***	0.518***	0.565***	0.576***	1.000									
	(0.000)	(0.000)	(0.000)	(0.000)										
(6) D_TotalBlockingTime	0.053***	0.048***	0.231***	0.442***	0.165***	1.000								
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)									
(7) Size Desktop	0.087***	0.223***	0.345***	0.432***	0.537***	0.179***	1.000							
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)								
(8) M_serverresponsetime	0.528***	0.143***	0.320***	0.094***	0.136***	-0.011	0.073***	1.000						
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.407)	(0.000)							
(9) M_firstcontentfulpaint	0.249***	0.781***	0.582***	0.382***	0.433***	0.066***	0.213***	0.197***	1.000					
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)						
(10) M_speedindex	0.335***	0.463***	0.692***	0.600***	0.533***	0.228***	0.394***	0.392***	0.651***	1.000				
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)					
(11) M_interactive	0.151***	0.382***	0.548***	0.755***	0.586***	0.323***	0.465***	0.109***	0.443***	0.771***	1.000			
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
(12) M_largestcontefulpaintt	0.159***	0.409***	0.480***	0.472***	0.644***	0.137***	0.418***	0.119***	0.491***	0.589***	0.604***	1.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			
(13) M_TotalBlockingtime	0.054***	0.070***	0.253***	0.491***	0.167***	0.563***	0.158***	0.005	0.081***	0.287***	0.410***	0.176***	1.000	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	-0.706	(0.000)	(0.000)	(0.000)	(0.000)		
(14) Size Mobile	0.087***	0.229***	0.343***	0.429***	0.526***	0.174***	0.906***	0.082***	0.228***	0.416***	0.491***	0.450***	0.179***	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	

Table B4: Experiment 1: Below Average Counts (Baseline)

Below Average		All			Treated			Controls	
Counts	Freq.	Percent	Cum.	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	111	13.65	13.65	52	12.81	12.81	59	14.5	14.5
1	168	20.66	34.32	85	20.94	33.74	83	20.39	34.89
2	140	17.22	51.54	72	17.73	51.48	68	16.71	51.6
3	162	19.93	71.46	88	21.67	73.15	74	18.18	69.78
4	171	21.03	92.5	80	19.7	92.86	91	22.36	92.14
5	61	7.5	100	29	7.14	100	32	7.86	100
Total	813	100		406	100		407	100	

Table B5: Experiment 2: Below Average Counts-Desktop (Baseline)

Below Average		All			Treated			Controls	
Counts	Freq.	Percent	Cum.	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	898	14.54	14.54	439	14.22	14.22	459	14.87	14.87
1	1,023	16.57	31.11	522	16.91	31.13	501	16.23	31.1
2	936	15.16	46.27	452	14.64	45.77	484	15.68	46.78
3	1,011	16.38	62.65	514	16.65	62.42	497	16.1	62.88
4	1,200	19.44	82.09	605	19.6	82.02	595	19.27	82.15
5	1,106	17.91	100	555	17.98	100	551	17.85	100
Total	6,174	100		3,087	100		3,087	100	

Table B6: Experiment 2: Below Average Counts-Mobile (Baseline)

Below Average		All			Treated			Controls	
Counts	Freq.	Percent	Cum.	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	1,085	17.57	17.57	543	17.59	17.59	542	17.56	17.56
1	1,063	17.22	34.79	526	17.04	34.63	537	17.4	34.95
2	833	13.49	48.28	425	13.77	48.4	408	13.22	48.17
3	891	14.43	62.71	455	14.74	63.14	436	14.12	62.29
4	1,238	20.05	82.77	603	19.53	82.67	635	20.57	82.86
5	1,064	17.23	100	535	17.33	100	529	17.14	100
Total	6,174	100		3,087	100		3,087	100	

Table B7: Experiment 1: T-test Statistics by Treatment Status (Balance test)

	Control	Treated	Difference
	(Sd)	(Sd)	(t-test)
First-Contentful-Paint	1.399	1.417	-0.0171
	(0.767)	(0.722)	(-0.32)
First-Meaningful-Paint	1.556	1.552	0.00445
	(0.877)	(0.786)	(0.07)
First-CPU-Idle	2.454	2.525	-0.0713
	(1.488)	(1.484)	(-0.67)
Time-to-Interactive	3.090	3.288	-0.197
	(2.286)	(2.042)	(-1.27)
Ranking Keywords	26.416	23.408	3.008
	(153.803)	(132.568)	(0.30)
Speed-Index	3.176	3.361	-0.185
	(2.164)	(2.088)	(-1.22)
Domain Authority	17.283	16.946	0.337
	(14.002)	(14.085)	(0.34)
Below Average Count	2.360	2.371	-0.0114
	(1.498)	(1.561)	(-0.11)
Page Views	6245.8	2467.1	3778.7
	(64724.89)	(11810.29)	(1.13)
N	406	407	

Table B8: Experiment 2: T-test Statistics by Treatment Status (Balance test)

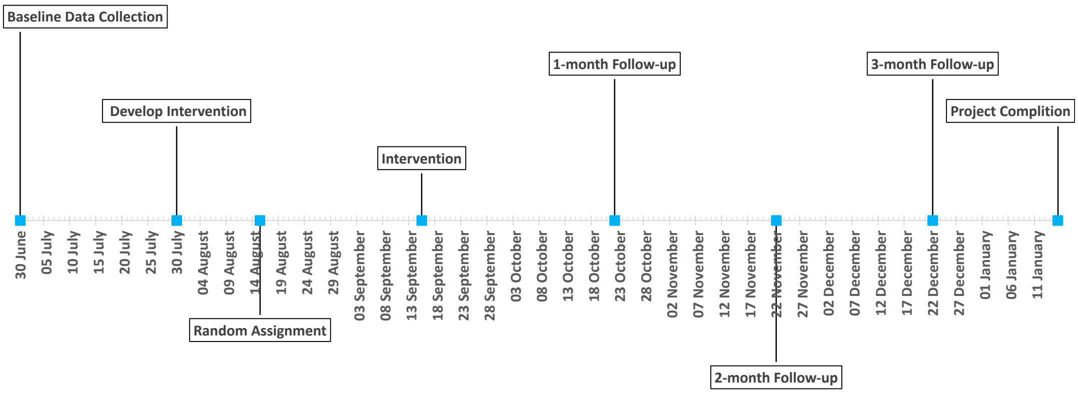
	(1)	(2)	(3)
	Control	Treated	Difference
	(sd)	(sd)	(t-test)
		Desktop	
Server Response Time	0.664	0.666	-0.00137
	(0.880)	(0.872)	(-0.06)
First Contentful Paint	1.005	0.999	0.00625
	(0.512)	(0.479)	(0.49)
First Meaningful Paint	1.110	1.103	0.00758
	(0.581)	(0.537)	(0.53)
Speed Index	2.314	2.312	0.00126
	(1.637)	(1.629)	(0.03)
Time to Interactive	2.190	2.206	-0.0160
	(1.818)	(1.790)	(-0.35)
Largest Contentful Paint	2.391	2.371	0.0197
	(1.830)	(1.771)	(0.43)
		Mobile	
Server Response Time	0.540	0.541	-0.00132
	(1.129)	(1.015)	(-0.05)
First Contentful Paint	3.979	3.957	0.0220
	(2.525)	(2.348)	(0.35)
First Meaningful Paint	4.575	4.595	-0.0196
	(2.748)	(2.663)	(-0.28)
Speed Index	8.424	8.499	-0.0752
	(5.724)	(6.525)	(-0.48)
Time to Interactive	12.21	12.30	-0.0812
	(8.689)	(9.487)	(-0.35)
Largest Contentful Paint	9.574	9.672	-0.0982
	(7.539)	(8.353)	(-0.48)
Observations	3087	3087	6174

Table B9: Experiment 2: T-test Statistics by Treatment Status and strata (Balance test)

Stratum	Without Analytic and Above			Without Analytic and Below			Without Analytic and Above			With Analytic and Below		
	Control	Treated	Difference	Control	Treated	Difference	Control	Treated	Difference	Control	Treated	Difference
	(sd)	(sd)	(t-test)	(sd)	(sd)	(t-test)	(sd)	(sd)	(t-test)	(sd)	(sd)	(t-test)
							Desktop					
Server Response Time	0.363	0.380	-0.017	1.033	1.039	-0.006	0.373	0.374	-0.001	0.933	0.926	0.008
	(0.379)	(0.425)	(-0.711)	(1.040)	(1.216)	(-0.074)	(0.372)	(0.371)	(-0.085)	(1.134)	(1.047)	(0.166)
First Contentful Paint	0.721	0.731	-0.011	1.401	1.329	0.072	0.787	0.782	0.000	1.194	1.202	-0.008
	(0.271)	(0.276)	(-0.662)	(0.711)	(0.644)	(1.505)	(0.210)	(0.219)	(0.540)	(0.532)	(0.491)	(-0.376)
First Meaningful Paint	0.762	0.773	-0.011	1.554	1.485	0.068	0.863	0.863	0.000	1.340	1.339	0.001
	(0.297)	(0.303)	(-0.603)	(0.790)	(0.733)	(1.272)	(0.234)	(0.242)	(0.016)	(0.604)	(0.532)	(0.044)
Speed Index	1.155	1.213	-0.058	3.412	3.316	0.096	1.397	1.374	0.022	3.289	3.311	-0.021
	(0.594)	(0.720)	(-1.505)	(1.709)	(1.767)	(0.783)	(0.612)	(0.584)	(0.817)	(1.710)	(1.680)	(-0.298)
Time to Interactive	1.046	1.019	0.027	2.821	2.785	0.037	1.321	1.294	0.027	3.283	3.378	-0.094
	(0.704)	(0.639)	(0.683)	(1.758)	(1.700)	(0.301)	(0.761)	(0.723)	(0.807)	(2.116)	(2.023)	(-1.089)
Largest Contentful Paint	1.239	1.287	-0.048	3.311	3.152	0.159	1.607	1.616	-0.009	3.315	3.287	0.028
	(0.658)	(0.666)	(-1.240)	(2.107)	(2.066)	(1.081)	(0.624)	(0.648)	(-0.312)	(2.144)	(2.060)	(0.315)
							Mobile					
Server Response Time	0.240	0.256	-0.016	0.944	0.881	0.063	0.277	0.279	-0.002	0.771	0.788	-0.017
	(0.416)	(0.386)	(-0.680)	(1.703)	(1.581)	(0.543)	(0.369)	(0.367)	(-0.115)	(1.417)	(1.232)	(-0.299)
First Contentful Paint	2.541	2.604	-0.063	6.142	5.620	0.523	2.741	2.793	-0.053	4.991	5.042	-0.052
	(1.049)	(1.110)	(-0.998)	(3.590)	(2.845)	(2.295)	(0.861)	(0.881)	(-1.326)	(2.544)	(2.560)	(-0.486)
First Meaningful Paint	2.874	2.916	-0.042	6.960	6.593	0.368	3.212	3.275	-0.062	5.747	5.860	-0.113
	(1.280)	(1.320)	(-0.552)	(3.679)	(3.204)	(1.516)	(1.132)	(1.123)	(-1.211)	(2.701)	(2.774)	(-0.982)
Speed Index	4.095	4.107	-0.012	12.435	12.147	0.288	4.962	4.985	-0.024	12.133	12.420	-0.287
	(2.114)	(2.122)	(-0.094)	(5.946)	(8.000)	(0.582)	(1.891)	(1.919)	(-0.270)	(5.635)	(6.850)	(-1.091)
Time to Interactive	5.632	5.255	0.377	15.690	15.455	0.235	7.847	7.774	0.074	18.033	18.599	-0.566
	(3.812)	(3.470)	(1.771)	(7.542)	(8.852)	(0.406)	(3.582)	(3.597)	(0.448)	(9.472)	(10.630)	(-1.341)
Largest Contentful Paint	4.974	4.824	0.150	13.693	12.971	0.722	5.962	6.043	-0.081	13.510	14.048	-0.538
	(3.102)	(2.777)	(0.873)	(8.486)	(7.718)	(1.266)	(2.689)	(2.744)	(-0.654)	(8.609)	(10.560)	(-1.333)
Observations	586	586	1172	405	404	809	957	958	1915	1139	1139	2278

The time line for the second experiment is available in Figure B1 below.

Figure B1: Time Line



C Mobile-Based Website Performance Metrics

C.1 Experiment 1: Overall Effect

Table C1: Mobile-Based Website Performance Metrics at Baseline

	count	mean	sd	min	max	p10	p25	p50	p75	p90
First-Contentful-Paint	785	5.294	3.027	0.600	24.3	2.6	3.4	4.5	6.6	9
First-Meaningful-Paint	785	5.898	3.451	0.600	28.2	2.9	3.6	5	7.2	10.1
Speed-Index	785	10.547	6.656	0.600	49.2	4.1	5.9	9.1	13.3	19.4
First-CPU-Idle	785	10.381	6.251	0.600	55	4.2	6.4	9.1	12.7	17.4
Time-to-Interactive	785	14.477	9.954	0.600	95.2	5.1	8.3	11.9	17.7	26.8
Observations	785									

Table C2: Regression Results for Mobile Speed (in Changes)

			Month One					Month Three					Month Six		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	ΔFCP	ΔFMP	$\Delta FCPUI$	ΔTTI	ΔSI	ΔFCP	ΔFMP	$\Delta FCPUI$	ΔTTI	ΔSI	ΔFCP	ΔFMP	$\Delta FCPUI$	ΔTTI	ΔSI
Treated	0.059	0.172	-0.114	-0.985**	-0.322	-0.116	0.071	-0.443	-1.674***	-0.955**	-0.102	0.062	-0.270	-1.239**	-0.495
	(0.138)	(0.155)	(0.338)	(0.490)	(0.316)	(0.148)	(0.168)	(0.359)	(0.559)	(0.373)	(0.174)	(0.198)	(0.452)	(0.589)	(0.427)
Cons	-0.672***	-0.924***	-0.388	-0.542	-1.104***	-0.805***	-1.155***	-1.180***	-1.011**	-1.435***	-0.901***	-1.259***	-3.846***	-0.894**	-0.550*
	(0.115)	(0.133)	(0.250)	(0.434)	(0.245)	(0.118)	(0.137)	(0.268)	(0.469)	(0.283)	(0.153)	(0.168)	(0.363)	(0.452)	(0.327)
R^2	0.005	0.007	0.002	0.007	0.004	0.008	0.005	0.003	0.016	0.012	0.005	0.005	0.003	0.014	0.007
N	758	758	758	758	758	763	763	763	763	763	761	761	761	761	761

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. FCP, FMP, FCPUI, TTI and SI are the five performance metrics that captures different aspect of page loading speed. They respectively stands for First-Contentful-Paint, First-Meaningful-Paint, First-CPU-Idle, Time-to-Interactive and Speed-Index. * Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

C.2 Experiment 2: Overall and Heterogeneity Analysis

Table C3: Experiment 2: Results for Mobile Devices, Changes by One, Two and Three Months

	(1)	(2)	(3)	(4)	(5)
	ΔSRT	ΔFCP	ΔTTI	ΔLCP	ΔSI
			Panel A: One Month		
Treated	0.000259	0.0493	-0.0333	-0.127	-0.136
	(0.028)	(0.040)	(0.166)	(0.174)	(0.130)
Cons	0.139***	-0.221***	-0.261**	0.0247	0.156*
	(0.022)	(0.029)	(0.114)	(0.121)	(0.092)
R^2	0.000229	0.00117	0.00141	0.000479	0.00117
N	5998	5998	5985	5989	5982
			Panel B: Two Months		
Treated	0.00782	0.0708*	0.0337	0.101	-0.0777
	(0.025)	(0.042)	(0.171)	(0.185)	(0.125)
Cons	0.0765***	-0.217***	-0.157	0.145	-0.242***
	(0.021)	(0.030)	(0.120)	(0.118)	(0.089)
R^2	0.000417	0.00112	0.00135	0.00106	0.000742
N	5855	5855	5855	5848	5852
			Panel C: Three Months		
Treated	0.0118	0.0819*	0.172	0.0118	0.0306
	(0.027)	(0.043)	(0.174)	(0.182)	(0.128)
Cons	0.0881***	-0.237***	-0.414***	0.117	-0.218**
	(0.021)	(0.031)	(0.120)	(0.124)	(0.089)
R^2	0.000191	0.00150	0.00121	0.00117	0.000768
N	5924	5924	5924	5919	5921

Notes: Robust standard errors in parentheses, and are clustered at firm level. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed (in seconds). They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index. * Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

Table C4: Experiment 2: Results for Mobile Devices, Changes by One, Two and Three Months Post-treatment by Below

	(1)	(2)	(3)	(4)	(5)
	ΔSRT	ΔFCP	ΔSI	ΔTTI	ΔLCP
			Panel A: One Month		
Treated	0.023	-0.050*	-0.018	0.046	-0.067
	(0.024)	(0.029)	(0.094)	(0.110)	(0.120)
BelowAvg	-0.252***	-0.985***	-0.972***	-1.167***	-1.680***
	(0.041)	(0.055)	(0.175)	(0.220)	(0.233)
Treated×BelowAvg	-0.046	0.206***	-0.240	-0.181	-0.126
	(0.054)	(0.077)	(0.256)	(0.325)	(0.338)
Cons	0.269***	0.272***	0.652***	0.333***	0.888***
	(0.016)	(0.022)	(0.066)	(0.083)	(0.096)
R^2	0.017	0.082	0.013	0.011	0.017
N	5998	5998	5982	5985	5989
			Panel B: Two Months		
Treated	0.017	-0.064**	0.006	0.090	0.077
	(0.017)	(0.030)	(0.087)	(0.139)	(0.157)
BelowAvg	-0.258***	-1.012***	-1.389***	-1.204***	-1.713***
	(0.039)	(0.056)	(0.166)	(0.234)	(0.229)
Treated×BelowAvg	-0.021	0.270***	-0.180	-0.138	0.039
	(0.049)	(0.081)	(0.244)	(0.334)	(0.361)
Cons	0.211***	0.290***	0.469***	0.459***	1.028***
	(0.012)	(0.023)	(0.062)	(0.101)	(0.111)
R^2	0.019	0.076	0.025	0.011	0.015
N	5855	5855	5852	5855	5848
			Panel C: Three Months		
Treated	-0.014	-0.028	0.013	0.088	0.104
	(0.021)	(0.031)	(0.096)	(0.132)	(0.149)
BelowAvg	-0.311***	-1.041***	-1.536***	-1.648***	-1.755***
	(0.040)	(0.057)	(0.167)	(0.229)	(0.234)
Treated×BelowAvg	0.049	0.225***	0.023	0.126	-0.190
	(0.053)	(0.082)	(0.249)	(0.337)	(0.353)
Cons	0.248***	0.284***	0.568***	0.430***	1.022***
	(0.016)	(0.023)	(0.071)	(0.100)	(0.096)
R^2	0.019	0.081	0.025	0.015	0.019
N	5924	5924	5921	5924	5919

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed. They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index.

* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

Table C5: Experiment 2: Results for Mobile Devices, Changes by One, Two and Three Months Post-treatment by Analytic

	(1)	(2)	(3)	(4)	(5)
	ΔSRT	ΔFCP	ΔSI	ΔTTI	ΔLCP
Panel A: One Month					
Treated	0.036	0.093	-0.292	-0.055	0.005
	(0.052)	(0.077)	(0.231)	(0.237)	(0.229)
Analytic	-0.018	-0.023	-0.263	0.012	0.130
	(0.045)	(0.066)	(0.188)	(0.219)	(0.232)
Treated×Analytic	-0.052	-0.064	0.227	0.032	-0.194
	(0.061)	(0.090)	(0.280)	(0.320)	(0.325)
Cons	0.151***	-0.205***	0.336**	-0.270	-0.064
	(0.038)	(0.058)	(0.155)	(0.165)	(0.174)
R^2	0.001	0.002	0.001	0.001	0.001
N	5998	5998	5982	5985	5989
Panel B: Two Months					
Treated	0.037	0.159**	-0.055	-0.098	0.287
	(0.049)	(0.079)	(0.232)	(0.266)	(0.265)
Analytic	-0.016	-0.005	-0.137	-0.108	0.058
	(0.044)	(0.068)	(0.187)	(0.246)	(0.236)
Treated×Analytic	-0.043	-0.128	-0.033	0.191	-0.270
	(0.057)	(0.094)	(0.276)	(0.342)	(0.358)
Cons	0.087**	-0.214***	-0.147	-0.082	0.105
	(0.037)	(0.059)	(0.159)	(0.200)	(0.185)
R^2	0.001	0.002	0.001	0.001	0.001
N	5855	5855	5852	5855	5848
Panel C: Three Months					
Treated	-0.009	0.117	-0.106	0.167	0.030
	(0.051)	(0.081)	(0.231)	(0.251)	(0.279)
Analytic	-0.037	-0.022	-0.101	-0.042	-0.078
	(0.046)	(0.070)	(0.184)	(0.233)	(0.243)
Treated×Analytic	0.031	-0.052	0.198	0.007	-0.027
	(0.061)	(0.096)	(0.278)	(0.337)	(0.363)
Cons	0.114***	-0.222***	-0.148	-0.385**	0.170
	(0.039)	(0.062)	(0.154)	(0.179)	(0.192)
R^2	0.000	0.002	0.001	0.001	0.001
N	5924	5924	5921	5924	5919

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. SRT, FCP, TTI, LCP, and SI are the performance metrics that captures different aspect of page loading speed. They respectively stands for Server-Response-time, First-Contentful-Paint, Time-To-Interactive, Largest-Contentful-Paint, and Speed-Index.

* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

D Experiment 1: Survey Results

In order to investigate how firms responded to the benchmark information and the reasons behind their responses, we sent a short survey to treated firms one year after sending the benchmark reports. As shown in Table D1, we received responses from only 27 firms. Among the thirteen questions included in the survey, we focus on the results from five, which exhibited the most apparent patterns across time.

Table D1: Summary Statistics of Survey Respondents

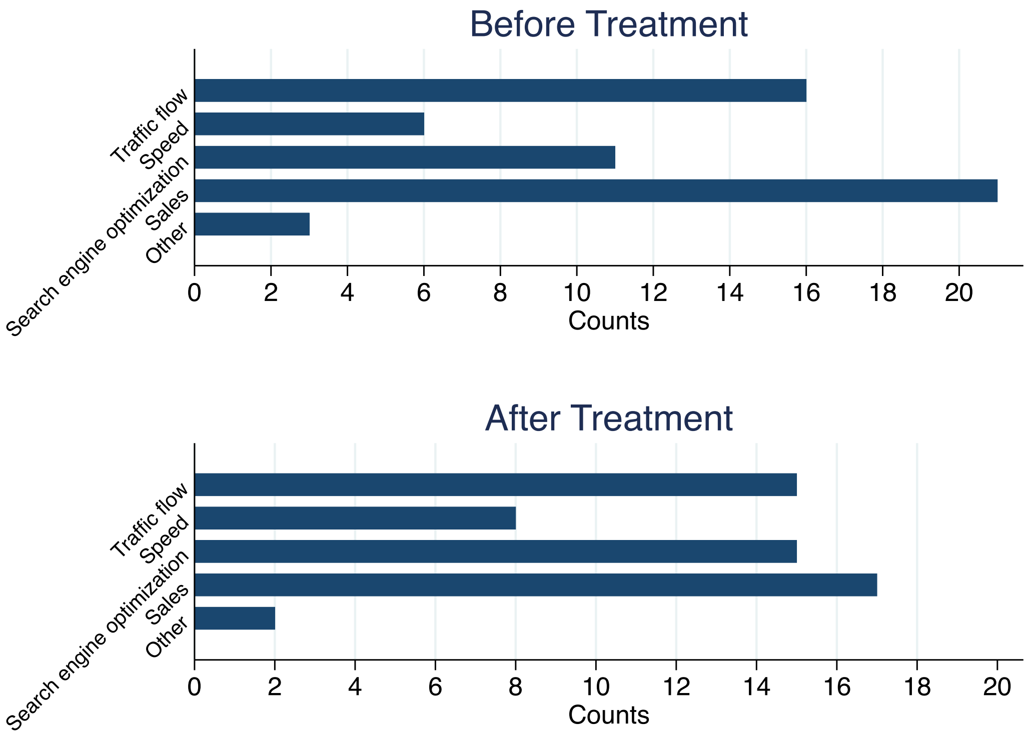
	count	mean	sd	min	max	p10	p25	p50	p75	p90
First-Contentful-Paint	26	1.665	0.904	0.700	4.5	.9	1	1.35	1.9	3
First-Meaningful-Paint	26	1.800	0.957	0.900	4.6	.9	1.1	1.45	2.3	3.3
First-CPU-Idle	26	2.781	1.153	0.900	5.2	1.6	1.9	2.6	3.4	4.7
Time-to-Interactive	26	3.569	1.815	0.900	8.3	1.7	2.2	3.1	4.7	5.9
Ranking Keywords	27	5.704	11.684	0.000	40	0	0	1	5	36
Speed-Index	26	3.858	2.497	0.900	9.6	1.6	2.4	2.9	4.9	8.7
Domain Authority	27	13.222	11.650	1.000	44	1	4	11	18	31
Observations	27									

Figure D1 presents the results from questions concerning the types of website performance metrics that were monitored at the start of 2021 (before the treatment occurred) compared to the post-treatment period. We allowed firms to select multiple metrics that were included in their monitoring processes. Prior to treatment, the results suggest that firms primarily monitored traffic flow and sales on their websites. Nineteen firms indicated they measured sales and 14 measured traffic flows. In contrast, only seven firms reported monitoring loading times, and ten monitored search engine optimisation. After the treatment, there is some evidence of increased usage of these additional performance measures. In the post-treatment period, the number of firms that monitored search engine optimisation metrics equalled those monitoring traffic flow (17), and the number of firms monitoring website loading times increased from six to eight.

The most noticeable impact is observed in the frequency with which firms monitored their website performance, as reported in Figure D2 and Figure D3. A clear trend toward increasing the monitoring frequency is evident when comparing pre-treatment and post-treatment periods. In the post-treatment period, there is an increase in the number of firms reporting daily, weekly, or monthly monitoring, and a decrease in the number of firms monitoring less frequently. Moreover, the number of firms stating they never monitor website performance dropped from six to three.

Lastly, we asked firms about the reasons behind their behaviour changes following the benchmarking treatment. As Figure D4 illustrates, half of the firms responding to the survey selected the option indicating "they were reminded of the need to take action." Other options were chosen much less frequently. The next most common response involved implementing actions to monitor performance (five firms) and using the benchmarks (four firms). Some firms also wrote to us with statements such as, "Since you sent me the last report, I have now found someone to look at my website and put the suggestions for improvements to performance into action" and "I hope to do something about the findings in it shortly." A few firms indicated they took no action, either because they already performed these actions (one firm), did not consider them important (three firms), or found they performed relatively well and subsequently took fewer actions (one firm).

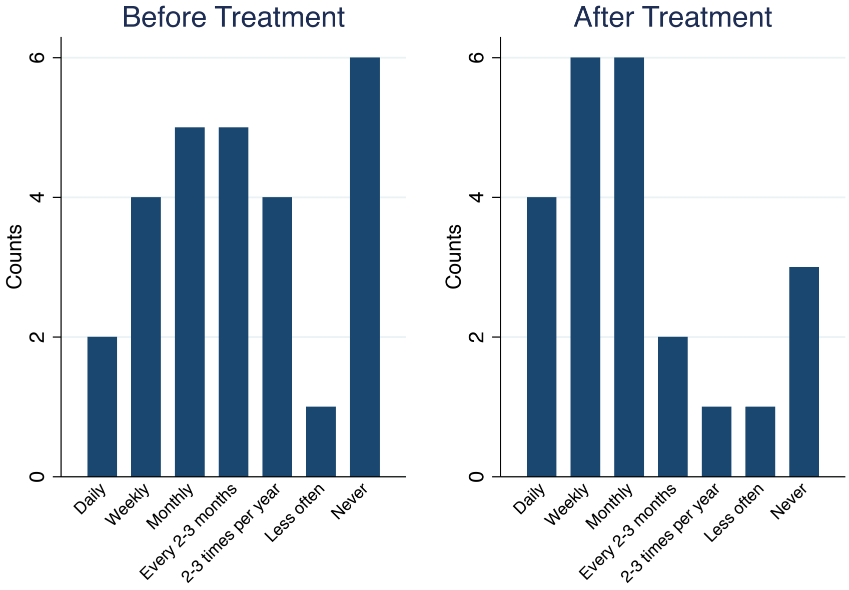
Figure D1: Types of website performance monitored



Notes: The survey asks them which dimensions of their website performance they monitor.

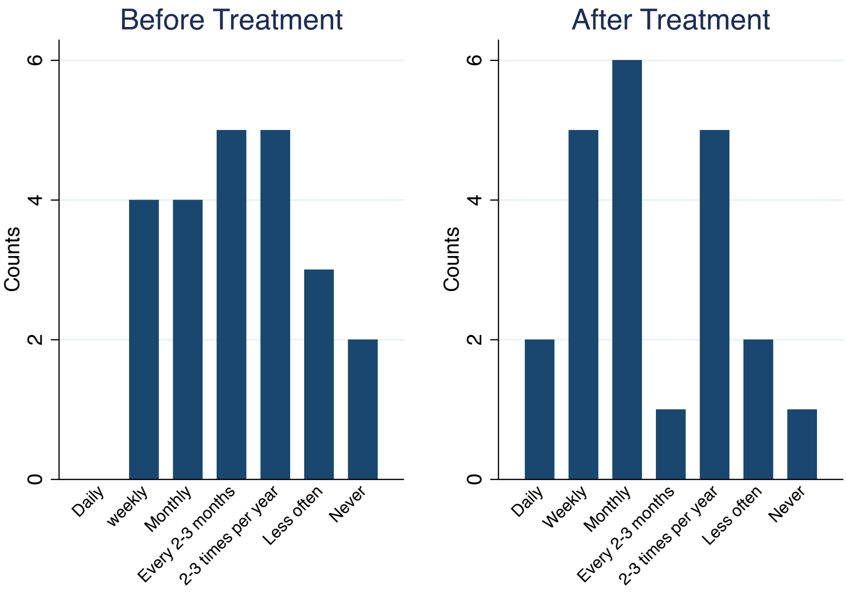
Overall, the results from the survey indicate that the benchmarking information treatment affected firm behaviours primarily by reminding them to take action on a more frequent basis, although there appear to be some firms that also put in place improved processes to monitor and improve the performance of their website.

Figure D2: Frequency of Information Collection on the Website



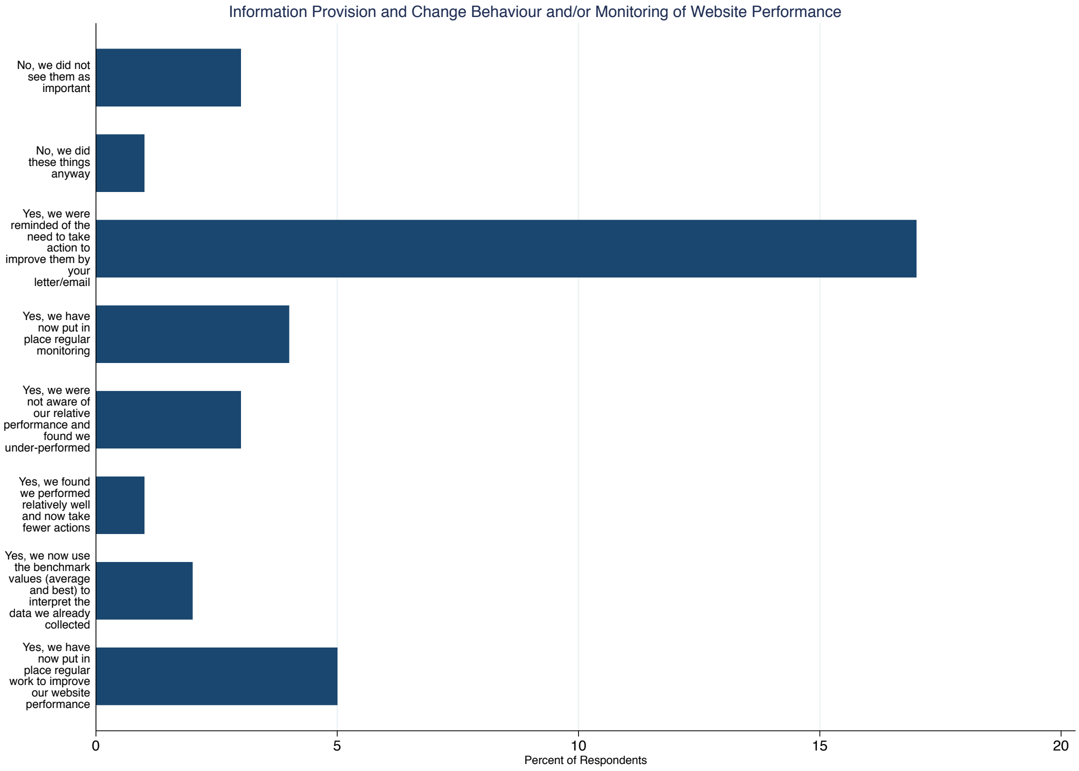
Notes: The survey asks them how often they collect information on their website performance.

Figure D3: Frequency of work to improve Website



Notes: The survey asks them how often they undertake work to improve their website performance.

Figure D4: Information Provision and Change Behaviour and/or Monitoring of Website Performance



Notes:Information Provision and Change Behaviour and/or Monitoring of Website Performance

E Experiment 1: Summary of Follow-up Experiment

To explore whether the results we generated were influenced by conducting the peer-benchmarking experiment during the Covid-19 lockdowns, when firms were heavily focused on their websites and their performance as a means of generating sales, we conducted a follow-up experiment on the firms that served as counterfactual in the first trial. For this experiment, we employed a staggered treatment design, labelled as early and later treatments.

The timeline for this experiment involved sending a letter to early treated firms at the beginning of the experiment window, informing them that they would receive peer-benchmarking information in one month, but not providing any data or additional information on how to improve website performance.[27]

[27] The decision to send this letter was informed by a result in the survey discussed in the subsequent section, which suggested

After one month, we send the early treated with information on their digital performance and benchmark, while late treated firms were sent a letter telling them they would receive their benchmarked data in one month. One month later the later-treated firms were sent the benchmarked data.

Table E1: Regression Results for One and Two Months in Changes

	(1) ΔFCP	(2) $\Delta TTFB$	(3) ΔLCP	(4) ΔDA	(5) ΔRK
			Panel A: One Month		
Early Treated	0.00602	-0.101	-0.129	-0.163	-9.235
	(0.025)	(0.109)	(0.124)	(0.188)	(16.394)
Later Treated	0.00997	0.0649	0.00362	0.0109	-28.36***
	(0.016)	(0.075)	(0.087)	(0.130)	(10.832)
R^2	0.00817	0.0162	0.0102	0.0563	0.0204
N	374	374	373	406	406
			Panel B: Two Months		
Early Treated	-0.00730	-0.0367	-0.193*	-0.0197	-9.235
	(0.033)	(0.092)	(0.105)	(0.198)	(16.394)
Later Treated	0.0172	-0.0636	0.0220	-0.0117	-28.36***
	(0.032)	(0.066)	(0.099)	(0.137)	(10.832)
R^2	0.00195	0.0202	0.0233	0.0449	0.0204
N	369	369	367	406	406
			Panel C: Three Months		
Early Treated	-0.0127	-0.115	-0.158	-0.134	-151.8
	(0.030)	(0.095)	(0.136)	(0.213)	(144.614)
Later Treated	0.0239	-0.0703	0.0516	0.104	-163.9***
	(0.023)	(0.063)	(0.098)	(0.155)	(54.306)
R^2	0.00676	0.0233	0.00426	0.0438	0.0210
N	364	364	363	406	406

Notes: Robust standard errors in parentheses, and estimations are clustered at firm level. Outcome variables are in changes and are measured in seconds. FCP, TTFBI and LCP are the three performance metrics that captures different aspect of page loading speed. They respectively stands for First-Contentful-Paint, Time-To-First-Byte and Largest-Contentful-Paint. Moreover, DA and RK are a search engine optimisation measures and represents Domain Authority and Ranking Keywords, respectively.
* Denotes significance at the 10% level, ** Denotes significance at the 5% level, *** Denotes significance at the 1% level.

The results are presented in Table E1. Before discussing the effects of the staged treatments, we note that the changes in website performance between the baseline and the month following treatment in Panel A are small. This contrasts with the changes observed in the main experiment, suggesting that the national Covid-19 lockdowns during this period may have encouraged firms to focus more attention

that firms responded in the first experiment because they were reminded to do so.

on their websites. The treatment effects in Panel A are similar to those found in Table 4, although the comparisons are complicated by changes in the metrics reported by the Google Lighthouse data used for data collection. The treatment effect on the FCP variable is small, equivalent to less than 0.01 of the baseline standard deviation (standard error is 0.025 standard deviations). The effects on TTFB and LCP are noticeably larger, 0.10 and 0.07 of the baseline standard deviation, respectively, but are imprecisely estimated (standard errors 0.11 and 0.06). The effects on domain authority and ranking keywords are also negative, though both less than 0.01 standard deviations.

In Panel B, we measure outcomes for the early and late treatment groups again relative to the baseline period. For the early treated firms, we observe little effect on the FCP loading time variable. The effect on TTFB has decreased to 0.04 of the baseline standard deviation, while the effect on LCP has increased to 0.10 (standard errors 0.06 and 0.05 standard deviations). We also find that the results for the later-treated group do not match those for the early-treated group in Panel A. As a reminder, the later treated group received a letter informing them of the forthcoming information benchmarking, similar to what occurred for the early treatment group in Panel A. In Panel B, we find that FCP and LCP slowed by 0.03 and 0.01 standard deviations, respectively, and are slightly more precisely estimated compared to the change between the baseline and month one (standard errors 0.06 and 0.05), while TTFB improved by 0.07 standard deviations (standard error 0.07).

We conclude from this that, despite the reduced power of this experiment, there is no strong evidence that firms responded significantly to their benchmark information.

Table E2: Descriptive Statistics at Baseline

	count	mean	sd	min	max	p10	p25	p50	p75	p90
First-Contentful-Paint	383	1.087	0.516	0.200	3.2	.6	.8	1	1.3	1.8
Time-to-First-Byte	383	0.858	0.921	0.040	5.34	.12	.28	.53	1.11	2.18
Largest-Contentful-Paint	383	2.632	1.929	0.200	20.1	1	1.6	2.2	3.1	4.4
Domain Authority	406	18.256	12.98	1.000	79	3	8	16	25	37
Ranking Keywords	406	291.7	1708	0.000	24100	0	4	28	118	316

Table E3: Balancing Test

	Late Treated Mean	Early Treated Mean	Difference (t-test)
		Page Loading Speed	
First Contentful Paint	1.112	1.061	0.0507
			(0.96)
Time to First Byte (TTFB)	0.849	0.867	-0.0178
			(-0.19)
Largest Contentful Paint	2.633	2.630	0.00349
			(0.02)
		SEO	
Domain Authority	18.16	18.35	-0.187
			(-0.15)
Ranking Keywords	228.7	354.6	-125.9
			(-0.74)
N	203	203	406

t statistics in parentheses

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

F Acknowledgements

The authors are grateful for the helpful comments received from Tim DeStefano, Chris Forman, Renaud Foucart, Sharique Hasan, Marit Hinnosaar, Toomas Hinnosaar, Hyunjin Kim, Rembrand Koning, Ramana Nanda, James Phipps, Phanish Puranam, and the participants of the 3rd Conference on Field Experiments in Strategy (CFXS), the 2022 Advances with Field Experiments (AFE) Conference (University of Chicago), the ZEW Conference on the Economics of ICTs, the PrOPEL Hub Workshop (University of Strathclyde), and the CRIBS Seminar (University of Nottingham). Zeyang Liu provided excellent research assistance. Errors and omissions remain our responsibility. This work was supported by the Economic and Social Research Council [Grant number ES/T001771/1]. AEA RCT Registry ID AEARCTR-0007125.

Digital Benchmark

COMPANY NAME

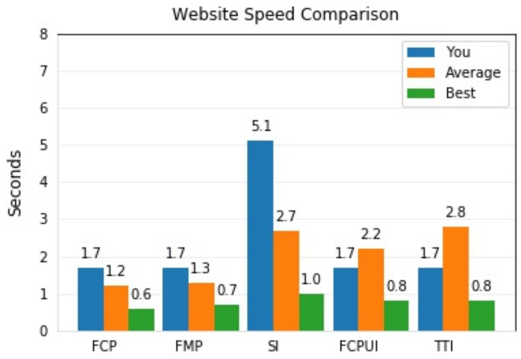
Researchers within the Schools of Business and Economics at the University of Nottingham have recently completed a comparison of website performance amongst UK distilling/brewing companies.

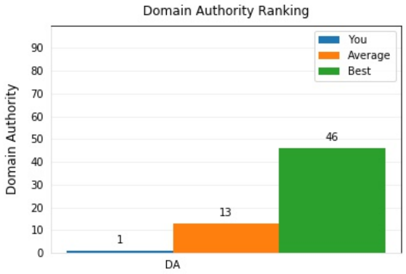
We can now share with your company management how you compare.

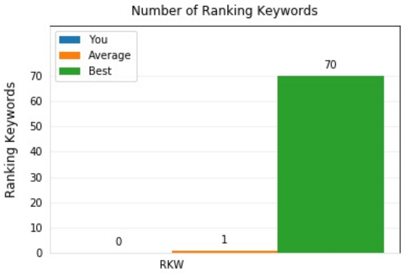
To make like-for-like comparisons we have benchmarked you against the best performing businesses with similar underlying web technologies to your own, as well as the average firm.

Website performance measures are created using data from batchspeed.com and moz.com.

This project is funded by the Economic and Social Research Council (ESRC) - a leading public funding body - to profile the use of digital technologies amongst UK companies and share what works.







Effective use of digital technologies is widely held to improve business competitiveness and productivity

Benchmark Indicators

First Contentful Paint (FCP): how long it takes the browser to load the first piece of content.

First Meaningful Paint (FMP): measures when the primary content of a page is visible to the user.

Speed Index (SI): measures how quickly content is visually displayed during page load.

First CPU Idle (FCPUI): measures how long it takes a page to become minimally interactive.

Time to Interactive (TTI): measures how long it takes a page to become fully interactive.

Domain Authority (DA): is a search engine ranking score developed by Moz.com that predicts how likely a website is to rank on search engine result pages (SERPs). A Domain Authority score ranges from 1 to 100, with higher scores corresponding to a greater ability to rank.

The Ranking Keywords: how many keywords your domain, subdomain or page is ranking.

Benchmark Indicators Matter

The indicators we have used to compare your website against are known to feed into Google and other similar search engine algorithms.

Data from Strange Loop showed that a 1-second delay in website loading time can reduce conversions by 7%, page views by 11% and customer satisfaction by 16%.

eConsultancy calculate 40% of people abandon a website that takes more than 3 seconds to load and 80% won't return.

Comparing gin/brewing businesses, for every 1 fewer keywords used, financial performance was 1% lower. For every 1 fewer linking root domain, it was 0.7% lower.

Ways to Improve

Increase website speed by

- Reduce the number of DNS (domain name system) lookups. For important webpages like your home page minimize external resources such as YouTube videos, a Twitter feed, Facebook page, and other content.
- Reduce the number of HTTP requests or use CSS Sprites or KeepAlive to make these act quicker.
- Use PHP Accelerators.
- Use GZIP compression to compress files.
- Reduce the number of redirects.
- Check for broken links on your pages. Fix or remove.
- Use browser caching.
- Reduce the size of images on your website.
- Use a Content Deliver Network (CDN) such as CloudFlare, EdgeCast, Cloudinary, CacheFly, MaxCDN

Improve linking route domains by linking out to websites with relevant content. This increases the value of your own website.

Keywords. Make sure pages and headings list product relevant words (gin, gin gifts, gin cocktails, gin&tonic, gin botanicals, gin glass, gin Christmas gifts, gin fizz...). Brainstorm everything that is relevant to what you do and where you do it - there are also keyword research tools that can help. However, remember it is humans who buy your products, not search bots.

Contact us

This information is provided as part of a knowledge exchange project to better understand website technology use, comparative website performance and business performance by UK companies.

To receive updates on your performance; for further information about the project; to opt out from receiving further information email: digitalbenchmark@nottingham.ac.uk

Project Leads:
Prof. Richard Kneller
Dr Cher Li
Dr Anwar Adem
Nottingham University Business School and School of Economics
propelhub.org