Journal of Quantitative Description: Digital Media 3(2023), 1–33 10.51685/jqd.2023.012

Testing, Testing: Identifying Contemporary Analytics Practices in Digital Politics

KATHERINE HAENSCHEN
CARL CILKE
ALISE BOAL
Northeastern University, USA

Digital analytics in contemporary politics receives tremendous attention from the media and has been the focus of a great deal of research over the last two decades. However, the actual practices that characterize work in the field often fail to receive sufficient attention. This paper presents the results of a quantitative content analysis describing the contents of 39 digital analytics case studies reporting the results of 68 individual A/B tests to learn about testing practices as they exist at high levels in contemporary U.S. politics. We find an emphasis on email and website testing, predominantly focused on fundraising and engagement outcomes. Our findings illuminate the mundane but substantive impacts of testing, which are predominantly focused on improving fundraising and email performance. Since firms made these case studies publicly available on their website, they also serve as marketing materials. In this manner we can understand how the practice of analytics is sold to political organizations looking to engage in digital testing.

Keywords: Digital analytics, A/B testing, political communication

Katherine Haenschen: katherine.haenschen@gmail.com
Date submitted: 2022-11-07

Digital technologies have changed the nature of organizing for political campaigns and advocacy groups. For many such organizations, the core functions of identifying people with relevant interests, communicating messages to them, and coordinating their contributions to the cause have largely moved online (Bimber, Flanagin, & Stohl, 2005). Digital analytics and the use of randomized controlled trials (RCTs) or so-called "A/B tests" in particular have emerged as the mechanism to optimize these online programs and measure successes, giving rise to what has been termed the "culture of testing" (Karpf, 2016; Baldwin-Philippi, 2016). Yet despite widespread attention on the existence of testing both within the academic literature and mainstream press (e.g., Baldwin-Philippi, 2017; Issenberg, 2012), we know little about how this practice is carried out day-to-day.

This paper presents a content analysis of 39 testing case studies made public by digital political consulting firms operating primarily or entirely in the United States.[1] Essentially, we quantitatively describe what is in a corpus of A/B testing case studies to learn about A/B testing practices; we do not conduct any experiments ourselves. By analyzing these case studies—the mediums used, variables manipulated and measured, analysis methods, results presentation, and firms who publish them—we can learn about testing practices as they exist at the high level of national political consulting firms that serve well-resourced candidates and may in turn generate findings that diffuse down to the local level. We choose to look at digital consulting firms because while most analytic advances occur at the level of presidential campaigns (see Kreiss, 2016), most testing *practice* takes place at firms such as these that employ a roster of analysts and work with numerous clients. By looking at trends and patterns across firms, we unearth evidence of what Karpf (2018, p. 4) terms "data-driven learning routines." In short, we can identify what is tested, how it is tested, and what kinds of causal inferences drive high-level digital political practice. Furthermore, since these firms make these case studies public on their

[1] We note that other types of analytics exist, namely observational studies that do not vary an independent variable nor assign treatment to randomized groups. While these analytics also have value, particularly for setting expectations in terms of outcomes such as email open or click-through rates or fundraising performance, our focus here is on analytics that use the experimental method.

website, they also serve as marketing materials, enabling us to explore how the practice of analytics is sold to political organizations looking to engage in digital testing.

We find 39 political case studies reporting 68 distinct experiments, all but one coming from left-leaning firms or firms that largely work with progressive or liberal organizations. Though Republican and right-leaning firms' websites were included in our search, they produced only one qualifying case study, reflecting prior work on the two parties' analytic professionalization. Surprisingly, the vast majority of named and anonymous clients in the corpus are advocacy and non-profit organizations, not electoral campaigns. The corpus shows an emphasis on optimization of email (38 tests, 55.9%), websites (16, 23.5%), and Facebook content or ads (14, 20.6%), with outcomes dominated by fundraising (25 tests, 36.8%) and email engagement (24, 35.3%). Results are largely presented as positive percent change improvements from one version to another. While the majority present descriptive statistics and reference statistical tests, actual specifics in terms of *p* values or sample sizes were infrequently included. We also identify language marketing firms' testing services to both external and internal clients.

Ultimately, our analysis leverages a corpus of practitioner texts to describe the important yet mundane practice of digital testing—dominated by fundraising and email performance—that characterize the practice of digital analytics by political consulting firms. These findings add depth and contextualization to our understanding of digital campaign practice and provide a helpful overview to scholars and students alike interested in digital testing practices. We discuss the implications of what we found, as well as what was noticeably missing from the corpus.

A Brief History of Digital Testing

Advances in digital testing practices in politics have largely been driven by U.S. presidential campaigns and national organizations, as most smaller entities lack the internal capacity and sample sizes necessary to conduct robust experimental programs.

One of the earliest organizations to adopt the testing methods of direct-mail fundraising for digital campaigning was MoveOn.org, a grassroots group that started in 1998 during the Clinton impeachment and gained further steam during the opposition to the Iraq War (Karpf, 2012). The group engaged in “exhaustive message-testing” while “generating millions in small-dollar contributions,” demonstrating the power of the Internet to raise funds and optimize this process through iterative experimentation (Karpf, 2012, p. 29).

By the 2004 presidential cycle, digital testing was practiced at the highest level of American politics. The Kerry campaign conducted A/B tests on their website to increase donations and email sign-ups, and the Dean campaign adopted email testing best-practices from MoveOn.org (Chadwick, 2007; Kreiss, 2012; Stromer-Galley, 2014). After the 2004 cycle ended, the groups Democracy for America (formed from the remnants of the Dean campaign) and MoveOn.org transitioned further into candidate and issue advocacy. With email membership lists in the millions, both groups adopted online testing practices and became incubators of analytics techniques and talent (Karpf, 2012). Dean alums launched Blue State Digital in 2004, which developed campaign technology and worked with the 2008 and 2012 Obama efforts (Kreiss, 2012).

By 2008, all of the presidential candidates had staff members dedicated to digital media (Stromer-Galley, 2014), with the Obama campaign investing heavily in "computational management" to make analytics-based decisions (Kreiss, 2012, p. 22). The use of large datasets by the Obama effort gave rise to so-called “data-driven campaigning,” which Baldwin-Philippi (2019, p. 2) defines as the use of “large data sets to either target messages to particular populations or test the efficacy of variations of messages and a variety of goals.” After 2008, Obama campaign alumni again began launching digital consulting firms, while workers with backgrounds in tech, marketing, and consulting left these industries to begin working in political analytics (Kreiss, 2012, 2016; Kreiss & Jasinski, 2016). For example, the firm Optimizely, which focuses on website testing, was started by Dan Siroker, who left Google to serve as Director of Analytics for the Obama

campaign (Kreiss, 2012; Siroker, 2010); the firm now provides testing services to a range of non-profit organizations.

As digital testing gained prominence at the national level, people and practices began to flow between organizations and campaigns, developing what Chadwick (2007) refers to as digital repertoires. By 2011, testing had become commonplace among national organizations, as way to make decisions grounded in data that reflected broader trends in political practice (Karpf, 2016). Testing practices largely converged into a hybridized form that combines the looseness of social movements with the traditional, tightly controlled structure of legacy organizations. Professional organizations on the left such as the New Organizing Institute, Rootscamp, and Netroots Nation began to provide training in digital analytics at annual gatherings, helping new staffers learn techniques needed to land jobs in this space (Baldwin-Philippi, 2016). Essentially, through the late 2000s and early 2010s digital analytics emerged as a professional field, with practitioners, firms, and—ideally—their clients dedicated to the culture of testing.

However, due to the partisan nature of politics in the United States, these practices essentially developed in parallel by Republicans and Democrats, with Democrats fielding a sizeable staffing advantage by 2016. Democrats invested more heavily in hiring individuals in positions related to technology, digital media, data, or analytics (Kreiss & Jasinski, 2016), with a particular interest in hiring people from outside of politics. Democratic professionals were also more likely to launch their own consulting firms and are credited with possessing the structural factors more likely to result in innovation (Kreiss & Jasinski, 2016; Kreiss & Saffer, 2017).

Here, a brief clarification of terms might be useful. In this paper, "testing" refers specifically to the use of the experimental method to conduct a randomized controlled trial (RCT). Subjects are randomly assigned to one of two or more groups, ensuring even distribution of covariates. The only difference is the treatment to which they are assigned, thus any difference in outcome must be attributable to the treatment condition. While many

RCTs in scientific research compare receipt of some treatment to a control group that receives nothing (or a placebo), digital testing often compares two versions of a treatment: Version A vs. Version B. As such, they are often referred to as A/B Tests, though in practice they are not limited to two variants.

A Broader Shift in Expertise

The increasing reliance on digital testing in politics is representative of a broader shift in terms of what kind of "expertise" matters in contemporary American corporate practices, particularly technology firms. In the early 2000's, staff at Amazon, Google, and Microsoft were already using A/B testing on various aspects of their websites (Christian, 2012; Kohavi et al., 2009). Since then, experimentally informed practices have become "almost a governing ethos" among Silicon Valley firms (Christian, 2012). Testing is often presented as the "correct" means of obtaining knowledge to make organizational decisions, and superior to relying on the HiPPO, or "Highest Paid Person's Opinion" (Kohavi et al, 2009). When an Amazon.com engineer developed the first personalized recommendation feature, executives were initially skeptical that it would boost sales. The engineer used an A/B test to demonstrate the boost in revenue, and based on the results executives immediately implemented it (Kohavi et al., 2009).

A simultaneous shift occurred in political campaign practices, driven by a series of academic field experiments demonstrating the effectiveness of tactics such as canvassing and phone-banking on voter turnout (e.g. Gerber & Green, 2000; Green et al., 2003). Consultants were also doing experiments on political mail as early as the 1990s and 2000s (C&E, 2012). These tactics sparked a sea change in campaign practices, as political professionals began to adopt experimentally informed tactics once they realized that the findings could help them win (Issenberg, 2012). In *Get Out the Vote*, Green and Gerber (2015) actively refute this notion that "experts know best" (p. 9), arguing that traditionally, consultants rarely measured the effects of their mobilization tactics, preferring to base decisions on "received wisdom."

Over time, testing has become part of conventional campaign wisdom, at least at the national level. On the left, Democratic and Labor groups formed the Analyst Institute in 2007, a proprietary clearinghouse of experimental results, which began compiling and conducting RCTs in voter mobilization (Issenberg, 2012). Republicans began doing field experiments in the early 200s as well, though by 2013 they recognized that they were trailing in these areas (C&E, 2012; RNC, 2013).

Digital testing is similarly posited as the antidote to fallible humans' inability to correctly predict a winner. One of the persistent stories from the 2008 Obama campaign describes the first splash page test conducted on the website. Staff members guessed incorrectly about which combination of image and text would win; they were wrong. The winning version increased sign-ups by 40%; had the staff gone with their guess, they would have lost out on millions of dollars in fundraising (Kreiss, 2012; Siroker, 2010). The same was true in 2012: Obama's director of digital analytics Amelia Showalter told *Bloomberg*, "We were so bad at predicting what would win that it only reinforced the need to constantly keep testing" (Green, 2012). Now, practitioners broadly emphasize the need to run testing programs: In *Campaigns & Elections*, a consultant states "I used to believe that, like good social media, email and website work was art. Now, it’s 100 percent science. ...no part of [a] digital campaign is left up to the opinions of anyone—everything is backed up with data" (Luidhardt, 2015). In this manner, testing has assumed the role of what Karpf (2016) calls the "neutral arbiter," able to settle disputes about messaging or email frequency.

Contemporary Political Digital Testing Practices

Most of what we know about actual testing practices comes from qualitative research—usually ethnographic or interview-based—that spends time with campaign professionals to understand what they do (e.g. Kreiss, 2012, 2016; Karpf, 2016; Baldwin-Philippi, 2016; Nielsen, 2011). This literature has helped identify the primary mediums for digital testing: email, organization websites, and the social media site Facebook. This work also highlights a range of other forms of observational analytic practices—website traffic

patterns, social media audience growth and engagement—that do not use RCTs. We acknowledge that this work exists and is valuable, but it is not the focus of our paper.

Despite the media attention paid to them, digital testing practices have not been adopted uniformly by political organizations (Baldwin-Philippi, 2016; Karpf, 2016). A large email list or volume of web traffic is needed to conduct tests with sufficient statistical power; niche or local organizations often lack this necessary and instead rely on results shared from larger organizations (Karpf, 2016). This need for scale as well as staff capacity has also limited many down-ballot campaigns from adopting analytical practices other than simple A/B testing on website and email content (Baldwin-Philippi, 2016). Below, we review specifics of testing practices by mediums in the academic literature, and then discuss what we know about their adoption, and how they contribute to theory-building.

Email Testing

Email is generally viewed as the most important digital tool for organizers, particularly given its role in fundraising (Gaynor & Gimpel, 2021; Nielsen, 2011). Email testing varies message and subject line content, as well as graphic design elements (Karpf, 2016; Kreiss, 2012; MacIntyre, 2020). “Campaigns’ email operations can measure how message elements like subject header, different content, layouts, or action buttons, effect the likelihood a recipient is to simply open the message, or take a subsequent action like donate money or sign up for an event” (Baldwin-Philippi, 2019, p. 4). For example, MoveOn.org’s found that adding a member’s ZIP code to an email subject line increased donation rates (Karpf, 2016). Organizations also test the "welcome series" of emails sent to new subscribers on downstream donating or unsubscribing (Kreiss, 2012).

Website Testing

Website testing largely focuses on which version of a layout or content produces the highest rate of new email sign-ups or donations, with particular emphasis on the

"splash" or landing page, and donation page (Kreiss, 2012; Stromer-Galley, 2014). Acquisition testing determines what kind of graphic design or layout inspires the greatest share of website visitors to sign up for the email list–funneling them into future testing in that medium (Kreiss, 2012).

Social Media Testing

Despite widespread attention to the use of social media platforms by political campaigns (e.g., Kim et al., 2019), the literature is thin in terms of specific ways in which tests are run. Most of the specifics refer to testing the performance of content (Kreiss et al., 2018; MacIntyre, 2020), though the outcome variables are often left unstated. Part of this may be due to a sort of in-house out-sourcing of tests to platform employees. As Kreiss and McGregor (2018) illustrate, Facebook sends staff to provide free consulting to campaigns on how to use their ad platform, which can obviate the need for campaign staff to learn how to do so themselves. The Trump campaign received widespread attention for the sheer volume of Facebook ad variants they claimed to have tested with the platform's help (e.g. Baldwin-Philippi, 2019; Kreiss & McGregor, 2018).

Notably, some practitioners refer in interviews to "tests" that may include statistical analysis but do not use the experimental method to collect data (Kreiss et al., 2018). True RCTs conducted on Facebook would primarily be done through advertising, as the platform offers a built-in split-test feature that makes it easy to run A/B tests varying either creative content, audiences, or placement (Facebook, n.d.). Twitter offers no built-in testing tool;[2] Instagram ads can also be A/B tested through Facebook's ads manager platform.

Building Practice, Not Theory

[2] See Coppock et al., (2016) for one of the few examples of a field experiment on Twitter conducted in partnership with a political organization.

Across this body of work, scholars repeatedly relate comments that analysts were not interested in developing theory about why something worked, but rather focused on finding the next tactic that would boost results. Karpf (2016) relates the anecdote about how MoveOn.org's use of ZIP codes in fundraising emails was adopted by other organizations until the seeming novelty wore off and it stopped boosting returns. As a practitioner explains (Kreiss et al., 2018, p. 12), "something that worked this month would completely just fail next month, and you'd have to find what the new thing is and just keep testing." This pattern appears in journalistic coverage as well, in which "[a]ttention to the amount of testing, rather than the substantive findings tests reveal or return on investment they yield is also common" (Baldwin-Philippi, 2020, p. 8).

Learning From Practitioner Literature

Other sources of evidence about testing practices exist: the so-called "gray literature" produced by political organizations and consulting firms detailing the results of individual experiments. Recall that most organizations and campaigns operating at the sub-national level lack the resources to conduct robust internal testing operations. Instead, campaigns with the financial means will hire digital firms specializing in this work.[3] The results of these tests form a literature that is largely proprietary and non-public, circulating on private listservs or stored in password-protected archives, such as that of the Analyst Institute (Issenberg, 2012; Karpf, 2012).

However, some digital consulting firms do choose to make selected case studies public on their website, providing insight into contemporary practices. Our research collects and analyzes these case studies to answer the following research questions: *What types of organizations–both firms and clients–are represented in this corpus? Through what platforms and mediums are tests conducted? What types of independent variables are manipulated? What types of outcomes are measured, and how? What kinds of results do*

[3] Other, less-resourced campaigns may simply try to implement test findings available to them through partisan groups or practitioner trainings, if they bother at all.

they generate, and how are results presented numerically? Furthermore, we also view these case studies as marketing documents, made public by these digital firms in an effort to promote their services to future clients; we also ask *how do these case studies market digital analytics to their current and future clients?* In this manner we can understand how the practice of analytics is sold internally to the field.

Methods and Materials

This paper presents a description of digital testing practices in contemporary politics, compiled from public case studies released by digital consulting firms. Below we report how we collected our sample and coded each case study, and how we perform our descriptive analysis.

Sampling

To compile our corpus of texts, we conducted two phases of internet searches, one in 2019 before this paper was first submitted for review and another in 2023 during the revision process to ensure no case studies were overlooked. We break this process out by each search phase.

Phase 1

In 2018, the first author created a list of digital political firms that were employed by or founded by staff from the 2008 and 2012 U.S. presidential campaign cycles (see Kreiss, 2016) as well as firms employed by presidential candidates during the 2016 cycle (Sticka, 2015) and press coverage of industry leaders (Wyler & LoGiurato, 2013), resulting in a list of 36 firms (see supplement for list).[4] In Fall 2019, undergraduate student research

[4] We chose to focus on U.S. firms because all authors study American politics and lack expertise to identify firms operating in other parts of the globe. We encourage scholars who study digital

assistants used Google's Advanced Search function[5] to search each firm's website for pages with relevant keywords ("test," "tests," "tested," "testing," "experiment," "experiments," "experimented," "experimenting", "trial", "trials", "RCT", "RCTs"), which generated 313 individual website pages. Each page was saved as a PDF to prevent data loss due to the ephemeral nature of web content.

In order to identify detailed A/B testing practices, we chose to focus on website pages consisting of specific case studies rather than pages referring generally to the fact that firms perform digital testing. In summer of 2021, the second and third authors coded each of the 313 pages for whether it was a case study. Inclusion was based on any of the following: The page referenced a specific client or clients (or an anonymous client); the page provided some detail about the work the firm did for the client; the page included any detail about research design or process; the page was categorized in a section of the firm's website referred to as "case studies" or similar ("client work"); the page gave very specific details about multiple similar programs run for multiple clients. Pages were deemed not to be case studies if they consisted of the biography of a staff person or available staff position, listed generic services provided by the firm, aggregated posts on the website (e.g., the page listing case studies is not a case study, but the pages linked to it are), or it is a "how to" list without specifics about a client or program (i.e. "how to conduct a split test on your website"). This resulted in 91 URLs with case studies, which were saved as PDFs.

Each case study was reviewed by the second and third author to determine whether any part of it used the experimental method. Case studies consisted of an experiment if they met any of the following criteria: clearly referenced the experimental method (randomization, a control and treatment group, and/or two treatment groups), or named an independent and dependent variable; referred to a "testing program," testing two versions against each other, or testing some aspect of a program; or used language such as "winner,"

politics in other nations to replicate a similar study to identify testing practices at the global level and facilitate cross-national comparisons.

[5] Available at https://www.google.com/advanced_search

"optimized," "performs better." Case studies were not experiments if they used language such as "put our new tool to the test" or "conducted an Election Day stress test." In cases of disagreement between coders, the first author broke ties. Krippendorff's α for nominal data was 0.806. This resulted in a list of 46 URL-PDFs with A/B testing case studies. Each PDF was coded for whether it reported multiple tests, defined at the level of design (e.g. one test uses email and another uses Facebook, or two separate message tests with different content in each). If multiple tests were found, coders added a row to the sheet so that each test's individual design could be coded separately. This resulted in 72 total A/B tests.

Phase 2

In 2023, while this article was under review and undergoing revisions,[6] authors were concerned that the corpus had become dated and sought to identify additional firms and tests. The first author performed a search of the FEC database for any payee in the year 2022 with "digital" in the disbursement description. This generated a list of 46,059 individual payments. The number of payments to each payee was calculated; payments to platforms such as Facebook and Google (for the placement of ads) were removed.[7] There were 1,810 total firms listed in the FEC records; however, of those, 600 only showed 1 payment. We focused on the 40 firms with the most payments received, which accounted for over half of all individual payments. We then performed the same Google Advance search listed above. This surfaced another 15 pages that were case studies that may consist of an A/B test; only 3 were deemed to be tests.

Variable Coding

In the first phase of coding, the second and third author coded each of the initial 313 URL-PDFs based on their content. Where necessary, intercoder reliability is reported

[6] The challenges caused by the COVID-19 pandemic resulted in substantial delays to this very labor-intensive project.
[7] See supplement for further details.

in the form of percent agreement, which is appropriate for two coders using nominal data.[8] Where coders disagreed, the first author broke ties after initial coding was concluded. In the second phase of coding, the first and third author coded all of the results data, as well as the three additional tests. This was done simultaneously over Zoom, with coding disagreements resolved in real time. Full coding instructions are available in the supplement. Variables are described below; frequencies are reported in the results section.

Client name: Recorded from the PDF if named, otherwise coded as anonymous.

Client type: Coders assigned client type by researching each entity's website. Categories consisted of partisan electoral campaign or organization (100% agr.), non-partisan electoral campaign or organization (97.2% agr.), advocacy group (84.7% agr.), private sector (97.2% agr.), and other (93.1% agr.). Given that our focus is on understanding practices in *political* testing, we eliminate tests done for non-political clients; as a result, nine case studies featuring private sector clients doing non-political marketing were removed as this analysis.

Client partisanship: Coded by the second author by looking up each named client's website and other references to their political and endorsement activity, potentially categorized as left-leaning, right-leaning, bipartisan (actively endorses/partners with both major U.S. parties), and neutral (avoids expressly endorsing members of either major party).

Test medium: Each individual experiment was coded by the medium in which the independent variable manipulation took place: email (88.9% agr.), website (87.3% agr.), Facebook ads or posts (95.2% agr.), Internet display ads (non-Facebook/non-video) (98.4% agr.), or other (90.5% agr.).

[8] Krippendorff's α often produces unreliable results when the coded variable is infrequent, even if percent agreement is high.

Independent variable: Coders categorized what aspect of the medium was manipulated. Responses consisted of textual content or messaging (93.7% agr.), fundraising amount (98.4% agr.), graphic design elements (92.1% agr.), or other (87.3% agr.).[9]

Test dependent variable: We grouped these broadly by purpose, categorizing them as fundraising (85.7% agr.), volunteering (98.4% agr.), email engagement (opens or clicks) (90.5% agr.), petition signatures (92.1% agr.), after-action sharing (98.4% agr.), and other/not specified (57% agr.).[10] Additionally, coders recorded exact terminology used to refer to how the outcome variable was measured. For instance, within a fundraising A/B test, the outcome for each test variant might consist of total amount raised, number of donors, revenue per fundraising email recipient, or average gift amount.

Finally, we consider how the results are presented. Coders recorded whether test outcomes were reported numerically using raw numbers, means, or percentages, or any other sort of descriptive statistic (87.3% agr.). Statistical testing was recorded if the case study referred to any sort of statistical testing or significance (81.0% agr.); coders also noted if the case study referred to a *p* value anywhere on the page, either in a table or the text itself (98.4% agr.), and if the study reported sample sizes (88.9% agr.) and mentioned any effort to analyze results by subgroup, i.e. donors vs. non-donors (98.4% agr.).

In phase 2 of coding, the first and third authors recorded how the results were presented numerically. We coded for the presence of the following: any results numbers (not words); results of any test variant; variant results presented as a percentage ("10%

[9] "Other" IVs: social sharing buttons; after-action sharing page version; use of welcome email; cross-promotion emails; remarketing campaign; use of digital ads. The IV is use of the tool, with groups randomized to receive it or not, so in these instances the medium is the IV as well.

[10] This was curiously low; upon further review and discussion between coders there were 9 studies with a final disposition of "Other" for any dependent variable. Most of these studies (n=5) *also* had one of the specific dependent variables (fundraising, email engagement, petition signatures, after-action sharing) included in the case study. Of the four that were coded as only having an "other" dependent variable, those dependent variables were policy support level, canvass response rate, online form conversions, and lead generation.

click-through rate on version A"); variant results presented as a raw number ("10,000 clicks on version A"); any numerical overall results; overall results as a raw number ("3,000 total clicks in the program"; "an overall increase in 50 donations"); overall results as a percentage ("a 10% increase in donations"), whether result was presented as a percent change ("10% improvement from version A to B") or a percentage point increase ("click-through rates increased by 2.2pp"); the numerical amount of the percent change; the numerical amount of the percentage point increase; any reported *p* values; and whether the results included a table, or separately a figure. Agreement was 100% since coding was done simultaneously over Zoom. We also identified marketing language in the case studies that suggested an implied audience of either future or existing clients, or attempted to in some way market the firm's practices.

This resulted in a final sample of 39 PDFs reporting the results of 68 individual A/B tests. We collected the year of each case study from a publication date stated in the text where present, or in the meta-data of the website code itself. Tests covered the years 2006 through 2022. There appears to be an uptick in case study publication immediately following the immense attention paid to the Obama 2012 re-election campaign's data operations. Otherwise, however, frequency of publication remains relatively flat, mostly ranging from 0 or 1 to 4 per year (Fig. 1).

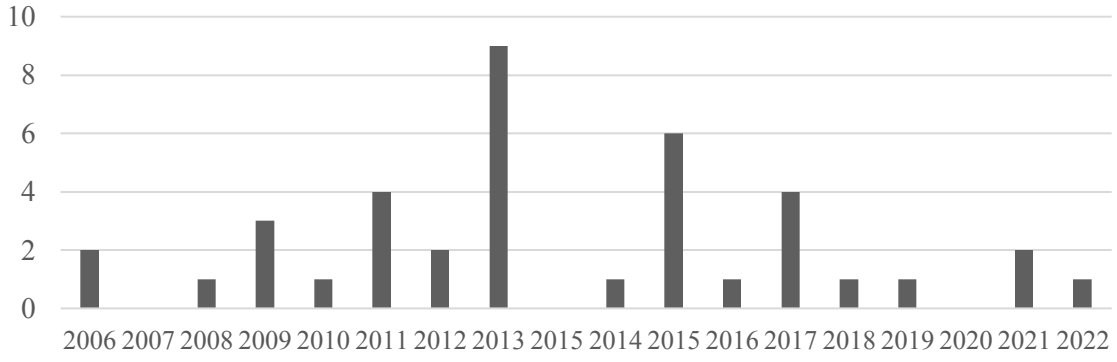

Figure 1. Frequency of case studies by year, 2006-present.

Analysis Method

We calculate frequencies and percentages to address our research questions and describe our dataset. Where appropriate we calculate crosstabulations between variables to understand how different experimental design features fit together.

Results

First, we report the firms and clients we find in our sample. Next, we describe the experimental designs: platforms, mediums, independent and dependent variables. Finally, we describe how results are presented numerically within the corpus.

Organizations

A search of 71 digital firms' web archives ultimately produced a list of seven firms with at least one political or advocacy digital A/B test case study retrievable from their website. The firms published a range of one (BlueLabs, IMGE, Trilogy Interactive) to 14 (ShareProgress) and 17 (M+R) separate case study URLs. Though the 71 digital firms that were searched contained 34 Republican-leaning firms, only one—IMGE—published any testing case studies. The remaining case studies came from either explicitly left-leaning firms (BlueLabs, GPS Impact, PowerThru Consulting, ShareProgress, Trilogy Interactive) or firms that work with non-profits that tend to lean left (M+R).[11]

We identified 23 distinct named clients within the test corpus, some of whom, such as the AFL-CIO, were featured in multiple tests. The majority of named clients were advocacy groups (20 of 23), followed by two partisan groups (Iowa Democratic Party, Progressives United), and one non-partisan campaign organization (For Our Future). Among the advocacy groups, we find national names in environmental issues (Sierra Club,

[11] Two other firms—a4 (f/k/a Audience Partners) and Precision Strategies—also do political work, but the case studies on their websites featured corporate clients, and were eliminated from analysis.

League of Conservation Voters, PETA, The Wilderness Society, Wildlife Conservation Society), human rights (Human Rights Campaign), civil rights (Color of Change), labor (AFL-CIO), and other charitable organizations (Easter Seals, Oxfam America, Save Darfur Coalition). The majority of clients (12) were identified as progressive or left-leaning organizations, followed by neutral (9) and bipartisan groups (2). None were explicitly right-leaning, as the right-leaning firm IMGE's client was not named.

Mediums and Variables

Next, we turn to test-level data, and the 68 distinct texts that make up our corpus. Since individual tests can contain more than one medium, we take a "check-all" approach; percentages may sum to more than 100%.

Test Mediums

Email was the dominant medium in the test corpus, amounting to 38 (55.9%) of tests reported. It was followed by website testing, 16 tests (23.5%); Facebook ad or content testing, 14 (20.6%); display ads, three tests (4.4%); and other, three tests (4.4%). Mediums grouped under "other" consisted of Twitter content in after-action shares, TV and digital ads, and video ads.

Independent Variables

Broadly speaking, the tests in our corpus focused on manipulating textual or message content (40, 58.8%), followed by graphic design (24, 35.3%), and fundraising amounts (9, 13.2%). Another nine tests manipulated elements grouped together as "other"—generally, these tests consisted of the use of a medium for treatment vs. receiving no treatment (see footnote 8 in methods section). We note that of the tests categorized as also varying an "Other" independent variable, all but two were also coded as varying messaging and/or design, since they tested whether receiving any of that type of

communication containing a specific message or design element had an impact related to receiving nothing.

This relationship is graphed below in Figure 2, with columns representing each medium, and segments of each column representing the independent variable, amounting to a total of 94 individual medium-IV combinations (remember, tests can and do include multiple mediums and multiple independent variables). We see a substantial amount of email tests pertaining to messaging—23, or 33.8% of all tests in our corpus. Next most common are design tests on websites (11, 16.2%), followed by email design tests (9, 13.2%) and message tests on Facebook (9, 13.2%).

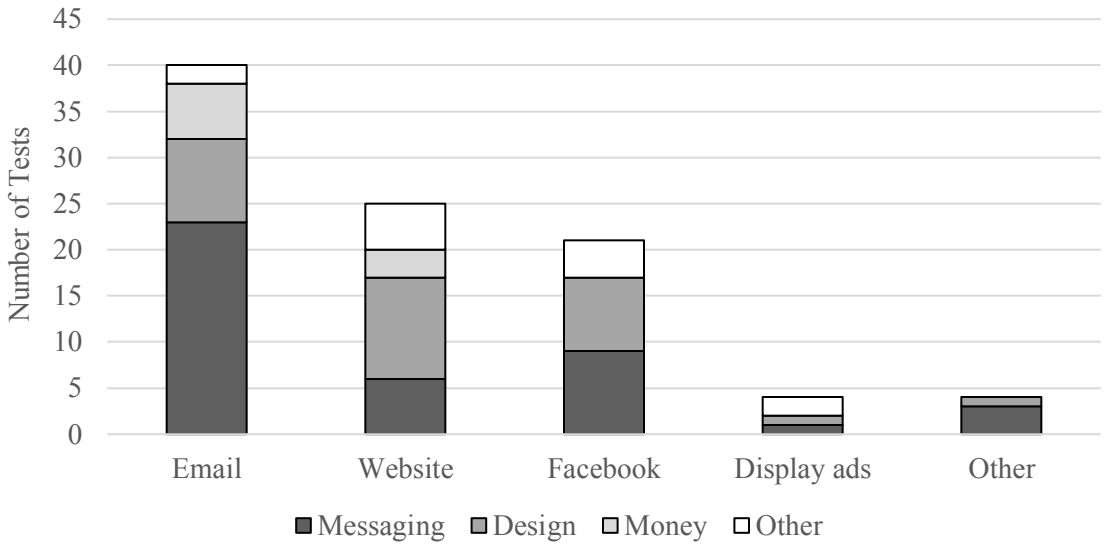

Figure 2. Frequency of test by medium and independent variable.

Dependent Variables

Turning to outcomes, we see that the plurality of tests in our corpus are focused on fundraising (25, 36.8%) and email engagement (24, 35.3%), followed by petition signatures (18, 26.5%). Tests measured in terms of after-action sharing—whether an

individual shares a petition via email or social media after signing—were well represented in the corpus (9 tests, 13.2%), primarily because a firm that specifically provides this service, ShareProgress, published 14 of the case studies we found. Tests coded as "Other" measured outcomes such as opinion change, canvass response rate, and online form submissions. Notably, though we looked for tests pertaining to volunteer activity, none of the case studies reported an experiment intended to increase such participation.

Within case studies pertaining to fundraising, independent variables are nearly evenly divided by varying the amount of money requested (9 tests), a design element (8), or the content of the textual appeal (7). Tests seeking to optimize engagement most commonly varied textual content such as subject lines (13 tests), followed by design elements (11). Every single petition test (18) varied the message content; two also varied design elements. We plot the frequencies of independent variables within dependent variables in Figure 3.

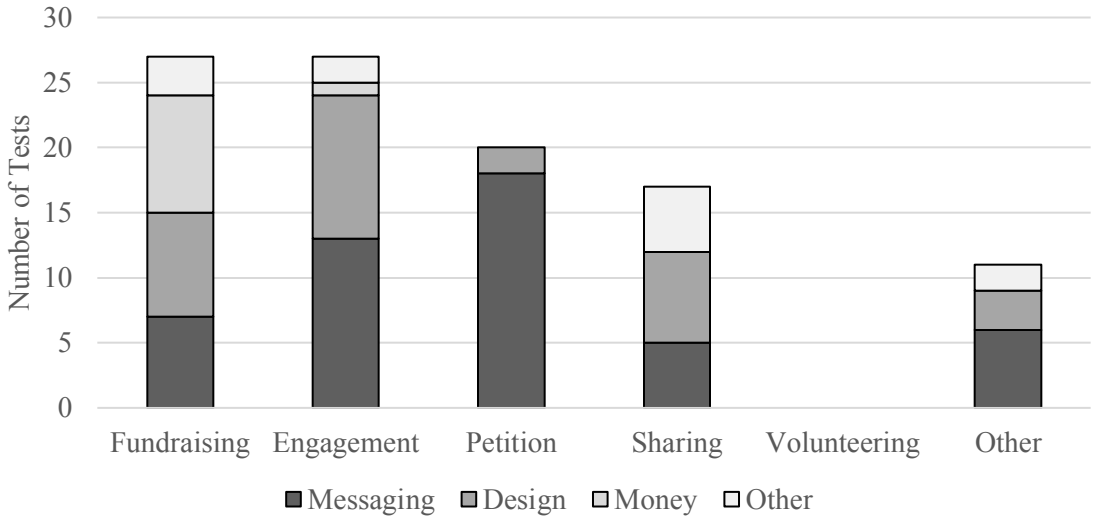

Figure 3. Frequency of test by dependent variable and independent variable.

Within our broad dependent variable categories we see ample heterogeneity in terms of how outcomes were measured. While fundraising tests as a whole were very common, individual tests varied in terms of how they measured success: total donations received, total revenue, average gift amount, and donor conversion rate were all common. Other tests calculated the ROI on digital advertising, such as cost per outcome. Email engagement outcomes were largely measured in terms of email opens, open rates, link clicks, click-through rates, total actions taken, and/or unsubscribes.

Results and Presentation Thereof

We next assess how the case studies present their results. Of the 68 total tests, 52 (76.5%) included some sort of numerical findings in the text. Another test did not report numerical results in the text but included a series of unlabeled charts. The remaining 15 (22.1%) tests did not offer any form of numerical results. A total of 24 tests reported results in large tables, often breaking out multiple dependent variables, such as click-through and unsubscribe rates for each test; eight tests included a figure.

Of tests that provided results, 39 presented overall tests results; 16 only or also presented some form of results for individual test variants. Overall esults were primarily presented in the form of percent change between variants (31 tests) rather than percentage point improvement (1 test). Other tests described overall performance (“29,000 shares”, “112,000 visits”, “24,000 new users”). The corpus of texts included 38 individual percent change effects values; these ranged from a 25.0% decline to 92.0% improvement. Overall, 30 effects were positive, 8 were negative. However, we note that effects can be positive in magnitude but negative in consequence, such as the test that found a 4.2% increase in people unsubscribing from the email list.

This decision to focus on percent change over percentage point improvement is strategic. For example, a test that improves an email open rate from 10% to 12.5% generates a 2.5 percentage point increase, and a 25% percent change. By presenting results

in this manner, firms are selling their results using the largest numbers available to them. Bigger is clearly better.

Turning to statistics, a clear majority of studies presented descriptive statistics of some sort (49, 72.1%), and a slightly smaller majority refer to statistical significance testing having occurred (44, 64.7%). Descriptive results and statistical testing co-occur: 54.4% of all tests have both; a significant chi-square test of independence [X^2 (1, N = 68) = 8.964, p < .01] suggests that this overlap is not random. However, case studies tend to be thin on statistical specifics. Only 11 (16.2%) reported a p value anywhere on the page; some case studies included more than one. Unsurprisingly, these scant p values tended to reach conventional levels of significance: two tests were marginally significant at the .10 level; three were significant at the .05 level; two at the .01 level, and six at the .001 level. Only nine (13.2%) included information about sample sizes. The lack of sample size may be a case of client privacy: advocacy groups don't want peers to know how large their email lists are. However, the lack of information makes any sort of formal meta-analytic estimates impossible. Only 2 (2.9%) tests included any sort of subgroup analysis, even though in practice many fundraising tests break out results for donors vs. non-donors.

Marketing Language

Finally, we consider the implied audiences for these tests, to determine to whom firm are selling their services. Surprisingly, most case studies appear to be intended for *existing* clients to make them aware of other services firms offer, perhaps in an effort to motivate clients to continue their monthly contracts. These case studies present findings in a neutral manner, often encouraging other organizations to replicate specific tests. Several invoke clients themselves: they feature interviews with client staff, or mention the firm's "amazing clients" and "brilliant campaign strategists" on the client side. This is in keeping with the placement of many case studies as posts on internal blogs: many firms' websites

appear to have now-defunct blog sections that were populated with case studies, news mentions, and other regularly recurring content.[12]

Only a handful of case studies seem explicit about bringing in new clients by exhorting readers to contact the firm. Phrases such as “Want to improve the open rates of your emails? Contact [firm] today!", ”You can put [our product] to the test with a 30-day free trial”, or "Let us know and we’re happy to work with you" are clear efforts to recruit new paying clients. Other case studies made a more nuanced sales pitch with phrases such as “hiring a team of experts to handle the nitty gritty details is a big investment – one that can pay off dividends.” The main feature that seemed to distinguish client recruitment from client maintenance was the use of more grandiose, hyperbolic language in the former: one firm “deployed quickly to meet the challenge”, another described test results as “remarkable”, and credited their work to a “program [that] has pretty much exploded”. These phrasings stand out from the more sedate, mundane tone of case studies that appear targeted to existing internal clients.

Discussion

Most of what we know about digital testing practices comes from qualitative researchers who spend extensive time with campaign professionals to understand what they do. Instead, this paper takes a quantitative approach: we set out to identify and describe testing practices as they exist in contemporary U.S. politics based on publicly available A/B tests reported by digital consulting firms. Much to our surprise, despite searching 71 firms’ websites, we found a small number of public case studies (39 individual URLs), the majority of which were from left-leaning firms promoting their work with non-profit and advocacy group clients. While we cannot content analyze what is not there, the inclusions

[12] We refer to these blogs as “now-defunct” because while our advanced Google search was able to find the pages on a firm’s website, the blog section itself is either no longer linked publicly in the main site navigation or is no longer being updated.

and exclusions are noteworthy in and of themselves. Below, we discuss the implications of what we found, as well as what was missing.

Describing Contemporary Digital Analytics Practices

While largely affirming prior work in terms of testing mediums, purpose, and partisanship, our findings bring new emphasis to the role of advocacy groups in sustaining analytic practices. Advocacy groups make for valuable paying clients: they engage in constant, year-round efforts to raise funds and engage members, and their need for ongoing services can translate into monthly fees that help firms keep the lights on. Conversely, electoral clients' activity may be concentrated around fundraising deadlines or voting periods, and many safe incumbents may spend no money on digital programs at all. Thus, the case studies primarily position firms to prospective non-profit organizations and advocacy groups that need to raise money online but may not be able to afford the overhead of an internal analytics team.

All but one firm with case studies in the corpus is left-leaning or works with liberal and progressive organizations; while there were non-partisan and bi-partisan groups represented, there were no right-leaning clients named in the corpus. This echoes prior work that finds an advantage among Democratic organizations stemming from the emergence of digital analytics practitioners in the 2004 cycle (Kreiss & Jasinski, 2016; Kreiss & Saffer, 2017). Notably, while Republican analytics firms do exist and we searched their sites for case studies, we found only one.

Descriptively, we find an emphasis on optimizing advocacy groups' core digital concerns: email engagement, fundraising, and list growth via petitions, echoing prior work (Karpf, 2016; Nielsen, 2011). This is done by manipulating primarily textual/message content or graphical design elements. These are relatively straightforward tests to conduct using existing features already built into tools commonly used to manage email lists and websites; there isn't much in the way of innovative research design here. Texts present

their numerical findings in the most positive light: percent change instead of percentage point increase, large increases in performance, *p* values but only when very small.

Collectively, these case studies position analytics consulting as an easy sell to clients–specifically to the decision-maker who needs to OK the expense. The case studies essentially say "hey, you're *already* raising funds, sending emails, and trying to engage members. We can help you do that even more effectively *and* measure it for you." In quantifying how political practitioners talk about their work to current and future clients, we distinguish these internal communications from how digital analytics is presented in the mainstream political press. Whereas the media—and practitioners who speak to reporters—often hype up analytic practice (see, e.g., Baldwin-Philippi, 2020), practitioners and firms present case studies in a relatively sedate manner.

Curious Absences in Our Corpus

The most notable absence in the corpus are tests involving electoral candidates, even at the statewide or federal level, as well as tests from right-leaning clients or firms. We know that nearly all Republican and Democratic candidates for federal office run some form of digital campaign involving email and social media pages (e.g., Macdonald et. al, 2022). Many of the digital firms whose websites were searched are known to do extensive work for candidates (e.g., Campaign Solutions, Mothership Strategies). Furthermore, based on our Google search results, most of these firms *do* run testing programs for clients: their websites emphasize the need to test, and job postings require experience with testing. However, most firms' sites offer no case studies about their A/B testing work with candidates.

So where are they? Candidates may not want their names associated with these practices, and firms want to keep outcomes internal as a form of competitive advantage. We also note that the vast majority of electoral candidates in the United States are seeking local office, and lack the means—such as an email universe offering suitable statistical

power (see Karpf, 2016 on list size; also Baldwin-Philippi, 2016) or staff with the knowledge—to conduct a testing program. The list size issue is likely also be true for some Congressional candidates. Yet we do not see any Gubernatorial, Senatorial, or Congressional candidate campaigns in our corpus—and we know from FEC data that these races are hiring analytics firms that engage in testing.

While we find references to statistical testing, statistical details such as sample sizes, p values and are largely lacking. We don't think that potential clients are choosing a digital firm based on whether their case studies acknowledge adequate statistical power. Quite the opposite–clients hire analytics firms to avoid thinking about these types of math problems! A more important question is whether these issues of statistical significance are engaged with by the analysts themselves and discussed with clients. Did a test "fail" because the sample was too small? Are all results reaching conventional ($p < .05$, or even $p < .10$ if there is reason to hold a directional hypotheses) levels of frequentist significance? Are clients actually receiving the *statistical* expertise they're paying for, and do they know enough to make sure?

Finally, we also note the lack of tests focused on fundraising through ads on Facebook or other digital media platforms, which is curious given their prominent role in the media. While practitioners state that they conduct Facebook ad tests (Kreiss et al., 2018), the firms in our sample evidently choose not to make these results public. This omission may be an effort to protect competitive advantage, as well as a desire to avoid scrutiny of their clients for engaging in practices that are indeed widespread in the industry but often receive outsized media attention.

Limitations of Our Sample

Our analysis is limited by what is in our sample, and while this project represents a substantial, multi-year effort to collect case studies, ultimately it can only speak for what can be found online, which in turn represents only what firms choose to publish. We make

no claims as to the representativeness of these findings as a stand-in for all digital analytic practice. The case studies that firms *do* choose to present offer insight into the practices and findings that they feel comfortable sharing publicly: email and website testing, optimization of engagement and fundraising, encouraging people to sign petitions and share them with friends.

We can thus infer from our corpus that firms have a strategic motivation to only present positive, non-controversial case studies on their websites. Other forms of analytics—such as microtargeted Facebook ads asking individuals to donate—are absent here, not because they do not happen, but because firms chose not to publicize them. Case studies tend not to report many negative or null results, instead positioning most results as positive.[13] We cannot tell from this corpus how many tests actually produce significant, meaningful increases in performance; neither can prospective clients using these texts to evaluate whom to hire.

Other Limitations, Next Steps

We only included U.S. firms because all three authors study American politics and are not sufficiently fluent in languages other than English; other scholars can and should replicate this analysis in regions where they have the expertise to identify firms working in digital politics. Though we searched the websites of 71 firms, including left- and right-leaning and bipartisan entities, we primarily found case studies from left-leaning firms; though right-leaning firms state on their websites that they do run testing programs. Based on these data, we cannot know if right-leaning testing practices vary.

We emphasize again that there are absolutely other examples of digital testing that circulate among practitioners, many of which are secured behind logins or paywalls or are kept proprietary within firms to retain a competitive advantage and justify clients' continued fees. Results might be different if those tests were able to be included, but these

[13] Evidently the file drawer problem exists for practitioners as well!

proprietary and hidden results are not available to the public, and researchers are unlikely to gain permission to analyze them for a broader audience.

Moving forward, qualitative researchers should use these descriptive findings—both the tactics present in them and those that are conspicuously absent—to inform future in-depth interviews with practitioners. Establishing descriptive quantitative benchmarks of this nature are an important first step to understanding contemporary digital analytics practices and how they may change over time, and the gaps we identified in the corpus should serve as fodder for future work.

Theory-Building vs. Practice-Building

We close by considering the general lack of theory-building presented in the documents. Generally, case studies do not surmise *why* something works, just that the A/B test was able to find the winning version. This focus on practice over theory both speaks to the challenges of temporal validity in digital research and is necessary given technological development and changes to the material reality of the Internet.

Karpf (2020) argues that the when and where of Internet practice matters as much as the what, since the material conditions of the Internet determine what the experience of using it—and conducting research through it—is like. However, he states that while the Internet changed rapidly throughout the 1990s and 2000s when campaign websites, email lists, and Facebook pages were first launching, the pace of technological breakthroughs slowed during the 2010s (Karpf, 2019b). For the last decade, the Internet has been dominated by the same major platforms, storage, and hardware providers (e.g. Facebook, Amazon, Apple, Google). This slowing rate of change provides "firmer ground for our most robust research methods" (Karpf, 2019b, p. 3)—in other words, doing an A/B test of an email subject line is a practice that can persist for a relatively long amount of time.

However, the results of any individual tactic ("does adding an individual's ZIP code impact open rates?") can and do change over time (Karpf, 2016). Munger (2022) refers to this as "temporal validity," arguing that most academic science is done with the goal of predicting the future; conversely, in digital analytics the goal of an individual A/B test is grounded in making a decision for the present moment, answering the question "what works better, right now?"

If list members' material experiences of the Internet change rapidly in terms of device, screen size, or email client, then digital testers should not necessarily expect the findings from an individual test to hold across multiple experiments, *if* the material and/or temporal conditions themselves are key to the results. And if something only works because it's a gimmick—ZIP codes in subject lines, yellow highlighting of donate links, graphic elements—over time that novelty effect will wear off.

The case studies we found suggest that digital analytics *practice* has solidified somewhat into an evergreen focus on email and fundraising (and especially email fundraising). Thus, it is knowing about *analytics practice*—what to randomly assign, how to measure the outcome—that matters in this field, more than the results of any individual test. Our corpus of texts reveals this: the digital repertoires and data-driven learning practices that comprise testing in contemporary digital political analytics.

Acknowledgements

The authors gratefully acknowledge and thank David Karpf for his feedback on this manuscript, as well as Alana Stillitano and Alex Springer for their work on data collection.

References

Baldwin-Philippi, J. (2016) The Cult(ure) of Analytics in 2014. In Hendricks J.A., Schill D. (Eds.), *Communication and Midterm Elections* (pp. 25-42). New York, NY: Palgrave Macmillan. doi: https://doi.org/10.1057/9781137488015_2

Baldwin-Philippi, J. (2019). Data campaigning: Between empirics and assumptions. *Internet Policy Review*, *8*(4), 1-18.

Baldwin-Philippi, J. (2020). Data ops, objectivity, and outsiders: Journalistic coverage of data campaigning. *Political Communication*, *37*(4), 468-487.

Baldwin-Philippi, J. (2017). The myths of data-driven campaigning. *Political Communication*, *34*(4), 627-633.

Bimber, B., Flanagin, A. & Stohl, C. (2005). Reconceptualizing collective action in the contemporary media environment. *Communication Theory, 15*, 365-388.

C&E. (2012, July 26). Shop Talk: The evolution of experimentation. *Campaigns & Elections.* Retrieved from https://campaignsandelections.com/industry-news/shop-talk-the-evolution-of-experimentation/

Chadwick, A. (2007). Digital network repertoires and organizational hybridity. *Political Communication*, *24*(3), 283-301.

Christian, B. (2012, April 25). The A/B test: Inside the technology that's changing the rules of business. *WIRED*. Retrieved from https://www.wired.com/2012/04/ff_abtesting/

Coppock, A., Guess, A., & Ternovski, J. (2016). When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior*, *38*(1), 105-128.

Facebook, n.d. About A/B testing, Retrieved from https://www.facebook.com/business/help/1738164643098669?id=445653312788501

Gaynor, S. W., & Gimpel, J. G. (2021). Small donor contributions in response to email outreach by a political campaign. *Journal of Political Marketing*, 1-25.

Gerber, A. S., & Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, *94*(3), 653-663.

Green, D. P., & Gerber, A. S. (2015). *Get out the vote: How to increase voter turnout*. Washington, D.C.: Brookings Institution Press.

Green, D. P., Gerber, A. S., & Nickerson, D. W. (2003). Getting out the vote in local elections: results from six door-to-door canvassing experiments. *Journal of Politics*, *65*(4), 1083-1096.

Green, J. (2012, November 29). The science behind those Obama campaign e-mails. *Bloomberg Businessweek.* Retrieved from https://www.bloomberg.com/news/articles/2012-11-29/the-science-behind-those-obama-campaign-e-mails

Issenberg, S. (2012). *The victory lab: The secret science of winning campaigns*. New York, NY: Broadway Books.

Karpf, D. (2018). Analytic activism and its limitations. *Social Media+ Society*, *4*(1), 2056305117750718.

Karpf, D. (2016). *Analytic activism: Digital listening and the new political strategy*. New York, NY: Oxford University Press.

Karpf, D. (2012). *The MoveOn effect: The unexpected transformation of American political advocacy*. New York, NY: Oxford University Press.

Karpf, D. (2019a). On digital disinformation and democratic myths. Social Science Research Council. Retrieved from https://mediawell.ssrc.org/expert-reflections/on-digital-disinformation-and-democratic-myths/

Karpf, D. (2019b). Something I no longer believe: Is Internet time slowing down?. *Social Media+ Society*, *5*(3), 2056305119849492.

Karpf, D. (2020). Two provocations for the study of digital politics in time. *Journal of Information Technology & Politics*, *17*(2), 87-96.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*, 140-181.

Kreiss, D. (2016). *Prototype politics: Technology-intensive campaigning and the data of democracy*. New York, NY: Oxford University Press.

Kreiss, D. (2012). *Taking our country back: The crafting of networked politics from Howard Dean to Barack Obama*. New York, NY: Oxford University Press.

Kreiss, D., & Jasinski, C. (2016). The tech industry meets presidential politics: Explaining the Democratic Party's technological advantage in electoral campaigning, 2004–2012. *Political Communication*, *33*(4), 544-562.

Kreiss, D., & McGregor, S. C. (2018). Technology firms shape political communication: The work of Microsoft, Facebook, Twitter, and Google with campaigns during the 2016 US presidential cycle. *Political Communication*, *35*(2), 155-177.

Kreiss, D., Lawrence, R. G., & McGregor, S. C. (2018). In their own words: Political practitioner accounts of candidates, audiences, affordances, genres, and timing in strategic social media use. *Political communication*, *35*(1), 8-31.

Kreiss, D., & Saffer, A. J. (2017). Networks and innovation in the production of communication: Explaining innovations in US electoral campaigning from 2004 to 2012. *Journal of Communication*, *67*(4), 521-544.

Luidhardt, K. (2015, April 8). 5 evolutions in digital. *Campaigns & Elections*. Retrieved from https://www.campaignsandelections.com/campaign-insider/5-evolutions-in-digital

Macdonald, M., Russell, A., & Hua, W. (2022). Negative Sentiment and Congressional Cue-Taking on Social Media. *PS: Political Science & Politics*, 1-6.

Macintyre, A. (2020). Adaption to data-driven practices in civil society organizations: A case study of Amnesty International. *Journal of Information Technology & Politics*, *17*(2), 161-173.

Munger, K. (2022, September 28). Temporal validity as meta-science. [OSF pre-print.] Retrieved from https://osf.io/hqkmr

Nielsen, R. K. (2011). Mundane internet tools, mobilizing practices, and the coproduction of citizenship in political campaigns. *New Media & Society*, *13*(5), 755-771.

RNC (Republican National Committee). (2013). The growth and opportunity project [Report]. Retrieved from https://www.gop.com/growth-and-opportunity-project/

Siroker, D. (2010). How Obama raised $60 million by running a simple experiment. *The Optimizely Blog: A/B Testing You'll Actually Use*, *29*.

Sticka, J. (2015, September 10). What can we learn from GOP FEC disclosures? Quite a bit, actually [Blog post]. Retrieved from http://www.risingtideinteractive.com/2015/09/gop-digital-2016/

Stromer-Galley, J. (2014). *Political discussion and deliberation online.* New York, NY: Oxford University Press.

Wyler, G. & LoGiurato, B. (2013). The digital 50: The 50 hottest people in online politics. *Business Insider*. Retrieved from http://www.businessinsider.com/digital-50-politics-tech-obama-republicans-2013-2

Supplemental Materials for "Testing, Testing: Identifying Contemporary Analytics Practices in Digital Politics"

KATHERINE HAENSCHEN
CARL CILKE
ALISE BOAL
Northeastern University, USA

Table Of Contents

I. Digital Consulting Firms

Table A1: Digital Consulting Firms Searched in Phase 1

Firm Name	Total Case Studies	Total A/B Tests
0ptimus	0	0
270 Strategies	0	0
Acquire Digital	0	0
ActionKit	0	0
ActionNetwork	0	0
Audience Partners (now a4)	0	0
Bask Digital Media	0	0
Blue State Digital	0	0
BlueLabs Analytics	1	1
Bully Pulpit Interactive	0	0
Campaign Solutions	0	0
Civis Analytics	0	0
ColdSpark	0	0
Direct Impact	0	0
Engage DC	0	0
gba strategies	0	0
Giles-Parscale	0	0
Harris Media	0	0
IMGE	0	0
Liz Mair	0	0
M+R	17	32
Mal Warwick Donor Digital	0	0
Mothership Strategies	0	0
Nation Builder	0	0
New Blue Interactive	0	0
Pantheon Analytics	0	0
PowerThru Consulting	3	3
Precision Strategies	0	0
Re:Power	0	0
Revolution Messaging	0	0
Salsa Labs	0	0
ShareProgress	14	25
Targeted Victory	0	0
Trilogy Interactive	1	1
Tusk Digital	0	0
Wellstone.org	0	0
Total	**36**	**37**

Note: This list consists of digital political consulting firms and digital activism platforms collected from academic and industry sources listing firms and individuals that worked on the 2008, 2012, and 2016 U.S. Presidential campaign cycles (Kreiss, 2016; Sticka, 2015; Wyler & LoGiurato, 2013). We list how many A/B test case studies and tests are in the corpus from each.

Table A2: Digital Consulting Firms Searched in Phase 2

Firm Name	Total Case Studies	Total A/B Tests
4 Degrees	0	0
Aisle 518 Strategies	0	0
Anne Lewis Strategies / MissionWired	0	0
Arena	0	0
Authentic	0	0
Base Engager	0	0
Battle Axe Digital	0	0
Blue Print Interactive	0	0
Break Something	0	0
Campaign Inbox	0	0
Campaign Solutions	0	0
Convert Digital	0	0
D-Ployit!	0	0
Donor Bureau	0	0
Fireside Campaigns	0	0
Go Big Media	0	0
GPS Impact	2	2
Hines Digital	0	0
IMGE	1	1
Mandate Media	0	0
Mothership Strategies	0	0
Middle Seat Consulting	0	0
New Blue Interactive	0	0
O2M Digital	0	0
Olympic Media	0	0
On Message	0	0
Push Digital	0	0
Reach Right Digital Marketing	0	0
Right Country Lists	0	0
Rising Tide Interactive	0	0
Run The World	0	0
Sapphire Strategies	0	0
Shiraz Media	0	0
Tag	0	0
Targeted Victory	0	0
The Prosper Group	0	0
Tma Direct	0	0
Veracity Media	0	0
Well & Lighthouse	0	0
West West Digital	0	0
Total	**3**	**3**

Note: This list was compiled by searching FEC records for all payments made in 2022 to any entity with "digital" in the description, and then the total A/B testing case studies and tests that were found through the search.

II. Coding Instructions

Below we report the coding instructions for each phase of data collection, categorization, and coding.

A. URL search using Google

1. Start with Google advanced search https://www.google.com/advanced_search

2. Type the firm URL into site or domain

3. Type keywords into "any of these words" and log number of URLs returned for each[1]

test OR tests OR tested OR testing OR
experiment OR experiments OR experimented OR experimenting OR
trial OR trials OR
RCT OR RCTs

4. Copy/paste each resulting URL into the "URLs" tab of the GoogleDoc

[1] Much to our grand annoyance, Boolean search operators such as * were not effective, such that test* did not return a full set of results for test / tests / tested / testing

B. URL coded as case study or not

1. Code each URL in the URL sheet for whether it is a case study using any method based on the criteria below. Mark each URL once you code it as Yes, No, Unsure.

Is this URL a case study?

Yes:	It references a single client, named or anonymous It provides some detail about what they did for the client It offers any details about research design or process It is in a section of the website called "Case Studies" or something similar ("client work") regardless of how thin / non-specific it is It gives very specific details about programs run for multiple clients
No:	It is a bio of a staff person It is a list generic services provided by the firm It is page that aggregates posts on the website It is a "how to" list without specifics about a client / program they ran
Unsure:	If you can't tell

2. If YES or UNSURE, save each case study as a PDF and upload to the folder using the page name as the file name.

C. Case study coded as A/B test or not

1. Open each numbered PDF. For each file, determine whether it uses the experimental method or not based on the criteria below.

Is this case study an experiment / A/B test / RCT?

Yes: It clearly describes randomization, control / treatment or 2+ treatment groups, and possibly any of the independent or dependent variables
It references a "testing program," "tested [something] against [something else]," or testing some aspect of digital media (website, email, Facebook, etc.)
It refers to multiple versions and something is a "winner" or "optimized" or "performs better" etc.

No: It uses "test" to refer to something other than what could reasonably be inferred to be an A/B test (i.e. "We conducted an Election Day stress test" or "put our new tool to the test")

Unsure: If you can't tell

D. Code each A/B Test for content

During Phase 1, two coders coded the entire corpus of A/B test PDFs, entering data into their own Google sheets. Here are the instructions for each column (each PDF is a row). The purpose of the coding was to mark each variable with a 1 or 0.

Column Letter and Name	**Instructions**
B: Coded by	Write your name
C: Organization / Firm	Which firm's website is this test from?
D: Multiple tests?	If a PDF contains one test, mark 0 and continue to column E. If a PDF contains multiple tests, mark Test 1, and add a new row below. Copy PDF number, Coder Name, and Organization to the new row. In that new row, mark Test 2. Repeat until each individual test has its own row.
E: A/B test	If the PDF / test is an experiment, mark 1 and move to column F. If the PDF / test is not an experiment, mark 0 and move to next row.
F: Client name	If the PDF / test names a specific client, mark 1 If not, mark 0 *Actual client names were collected by second author after initial coding*
G-L: Client type	In each column, mark if the client fits that description (1) or not (0) G: Partisan electoral campaign or organization H: Non-partisan electoral campaign or organization

	I: Political advocacy group J: Private sector / non-political K: Not specified L: Other, does not fit any of these categories
M-Q: Test medium	In each column, mark if the test uses that medium (1) or not (0). You may mark more than one column if multiple mediums are used. M: Website N: Email O: Facebook ads or posts P: Display ads not on Facebook Q: Other
R-U: Independent variable	What was manipulated in the test? Mark a (1) if that element was manipulated, (0) if not. Multiple elements can be manipulated. R: Messaging (words / verbal content) S: Money (donation amounts) T: Design (images, buttons, font sizes, image sizes) U: Other
V-AA: Dependent variable	What was the outcome? If the test measures a specific outcome, mark (1). If not, mark (0). Tests may measure multiple outcomes. V: Fundraising (outcome is any form of money coming in) W: Volunteering X: Email Engagement (opens, clicks, unsubscribes, etc.) Y: Petition Signatures Z: After-Action Sharing (if the test involves doing something after signing a petition or giving a donation, such as telling friends about it) AA: Other
AB: DV measured how	Copy and paste the exact wording used to describe how the dependent variable was measured.
AC: Descriptive statistics	Does the test refer to descriptive statistics in any way, or include descriptive statistics (“a 12% increase in X”)? If yes, mark (1), if no (0).
AD: Statistical testing	Does the PDF refer to statistical testing having occurred, or statistical significance? If yes, mark (1), if no (0).
AE: p value	Is there a p value explicitly stated in the test (including in a table)? If yes, mark (1), if no (0).
AF: Sample size	Does the test mention sample size? If yes, mark (1), if no (0).
AG: Subgroups	Does the test mention any type of subgroup analysis, e.g. results for donors vs. non-donors? If yes, mark (1), if no (0).

Additional coding was done during manuscript revision. Here are the instructions for each column (each PDF is a row). The purpose of the coding was to mark each variable with a numeric value, or a binary indicator, either 1 (yes) or 0 (no).

Column Letter and Name	**Instructions**
AH: Year	What year was the case study published? If not visible in text, look in HTML source code. Numeric.
AI: Source of Year	Was year in website text or HTML code?
AJ: Any results numbers	Are there any numeric results recorded anywhere in the text? (binary)
AK: Any variant results	Are there any numbers that refer to how one or more specific test variant(s) performed?
AL: Variant results percent	Are variant results reported using a percentage?
AM: Variant results raw	Are variant results reported using a raw number?
AN: Any overall results	Are there any numbers that refer to the overall performance of the test?
AO: Overall results raw	Record any raw numbers reporting the overall performance of the test
AP: Results percent change	Record any numbers reporting percent change between test variants
AQ: Results percentage pt	Record any numbers reporting percentage point increase between variants
AR: p value	Record all p values stated in test
AS: Table	Is there a results table anywhere in the test?
AT: Figure	Is there a results figure anywhere in the test?