COGITATIO

Media and Communication (ISSN: 2183–2439)
2019, Volume 7, Issue 1, Pages 117–127
DOI: 10.17645/mac.v7i1.1801

Article

Optimizing Content with A/B Headline Testing: Changing Newsroom Practices

Nick Hagar * and Nicholas Diakopoulos

Department of Communication Studies, Northwestern University, Evanston, IL 60201, USA;
E-Mails: nhagar@u.northwestern.edu (N.H.), nad@northwestern.edu (N.D.)

* Corresponding author

Submitted: 29 October 2018 | Accepted: 18 December 2018 | Published: 19 February 2019

Abstract
Audience analytics are an increasingly essential part of the modern newsroom as publishers seek to maximize the reach and commercial potential of their content. On top of a wealth of audience data collected, algorithmic approaches can then be applied with an eye towards predicting and optimizing the performance of content based on historical patterns. This work focuses specifically on content optimization practices surrounding the use of A/B headline testing in newsrooms. Using such approaches, digital newsrooms might audience-test as many as a dozen headlines per article, collecting data that allows an optimization algorithm to converge on the headline that is best with respect to some metric, such as the click-through rate. This article presents the results of an interview study which illuminate the ways in which A/B testing algorithms are changing workflow and headline writing practices, as well as the social dynamics shaping this process and its implementation within US newsrooms.

Keywords
audience metrics; content optimization; digital media; headline testing; headlines

Issue
This article is part of the issue "Emerging Technologies in Journalism and Media: International Perspectives on Their Nature and Impact", edited by John Pavlik (Rutgers University, USA).

1. Introduction

To stay in business, digital publishers depend on capturing the attention of users. Platforms like Facebook and Twitter use algorithms to surface and curate relevant content that drives user engagement. While often not as sophisticated, news organizations have also started incorporating data and algorithmic systems into their editorial workflows to optimize stories and capture reader attention. Such approaches are used in a variety of ways to optimize attention and traffic, including predicting article shelf-life, selecting and timing postings to social channels, and integrating recommendation and personalization modules to make sites more sticky. The integration of data and technology reveals new ways in which journalists respond to reader preferences. Audience influence has long been a factor in shaping journalism: newspapers used readership research to decide where they should increase coverage, and journalists received direct feedback in the form of mail and phone calls (Beam, 1995; de Sola Pool & Shulman, 1959). With the shift to digital, the task of constructing an audience has become increasingly quantitative, with analytics systems collecting feedback in the form of data (Zamith, 2018).

This article examines a specific technical approach that shapes the optimization of content by way of audience feedback: A/B headline testing. In particular, this research examines how headline testing is being taken up by news organizations and what its impacts are on headline writing. While many analytics and optimization tools influence journalism, headline testing plays a central role in shaping story presentation. By focusing on headline testing as a newsroom process, we illuminate how technology shapes a key component of publishing

and how different actors orchestrate its impact as part of a sociotechnical system. The crucial function headlines play in attracting attention to online news and the number of different newsroom roles that touch the testing process make headline testing well-suited to observing how technical, organizational, and editorial dynamics interact in media organizations. In this study, we offer findings from one part of those interactions, that of the newsroom workers who oversee and implement headline testing. By focusing on this group, we uncover an important perspective on an understudied practice.

To conduct our study, we undertook semi-structured interviews with 10 media industry professionals occupying a range of roles in headline testing and optimization processes. Using a grounded theory approach to analyze transcribed interviews, we develop a conceptual treatment of the social context of A/B headline testing within newsrooms and its role in shaping news headline writing. We find that newsroom workers frame A/B testing as a way to pursue "better" headlines by discovering and reinforcing headline writing best practices, to the extent that publishing contexts and the testing tool allow. We also detail how the social dynamics of the newsroom—which depend on established newsroom roles, productive communication, and credible influence—affect the headline testing process. In our discussion, we incorporate an actor-network theory (ANT) frame for examining the interplay between technology and newsroom workers exposed in our results. ANT considers the relationships among human actors, technical actants, and other objects, and the behaviors that emerge from those relationships, offering a way to understand the relationships between the human and non-human influences both within and outside the newsroom (Lewis & Westlund, 2015). We reflect on the mediating role of actors in the sociotechnical system, the implications of organizational scale for adoption, and discuss limitations and opportunities for future work in this domain. Taken together, these findings advance understanding of the newsroom dynamics surrounding the adoption and integration of headline optimization into newsroom practices.

2. Related Work

As digital media has evolved out of print practices, we consider two key changes that have affected how newsroom workers package and distribute their stories. First, audience analytics exert influence on journalists and their priorities. Second, tools designed to distribute or monitor content have changed workflows in the newsroom. By examining how newsroom workers are using new tools that track audience data to optimize headlines, this research sits at the intersection of these trends.

2.1. Audience Analytics

The growing influence of technology on journalism has shifted the relationship between the press and its audience, from unilateral, asymmetrical communication to "a dialog between the press and the public" (Pavlik, 2000, p. 235). With that dialogue have come new forms of incorporating audience feedback. Without the audience's active participation, journalists can indirectly observe and make inferences about reader behavior, extracting patterns of readership to inform publishing decisions (MacGregor, 2007). Readers generate user data by their actions online, making them meaningfully measurable (Assmann & Diakopoulos, 2017). Their data now informs many aspects of assigning and distributing stories, including shaping headlines (Jenner & Tandoc, 2013). That data manifests in the form of common metrics, often conveyed through third-party analytics services. Metrics like pageviews and engaged time serve as currencies, assigning value to interactions and transforming the industry around them (Nelson & Webster, 2016). The impact of those currencies depends on how central a newsroom makes audience analytics, in terms of visibility, prioritization, and utilization in the journalistic process (Petre, 2015). It also depends on structural factors—such as analytics systems, the sophistication of which vary by newsroom—and the newsroom workers whose labor interacts within that structure (Cohen, 2018; Nielsen & Cherubini, 2016).

While the tools for collecting audience feedback have changed, concerns about incorporating that data into the journalistic process remain. Critics of newspaper audience research saw it as a way to "pander to readers," rather than focusing on the information they needed (Beam, 1995, p. 28). More direct audience involvement in producing stories has spurred new anxieties. Tandoc and Thomas (2015) warn that web analytics can segment audiences into ever-narrower groups, eliminating the common ground where civic discourse takes place. There is also evidence that journalists' news judgment is eroding as a driving force behind production and distribution: Lee, Lewis and Powers (2014) found that audience data decides how prominently stories get placed on homepages more than editorial judgment does. And in some cases, reader data is used without the mediation of any editorial judgment, as in the case of "most read" lists (Lee et al., 2014). This shift can cause tension between the stated, traditional values of a newsroom and the pressure to incorporate reader data (Anderson, 2011). Metrics can also have a deep emotional effect on journalists and can decouple the goal of producing traffic from that of writing a high-quality story (Blanchett Neheli, 2018; Nelson & Tandoc, 2018). As Nielsen and Cherubini (2016) acknowledge, important signifiers of success and quality, like the public impact of a piece of reporting, won't be reflected by editorial analytics. This research aims to interrogate these tensions in the context of newsroom A/B headline testing, which tightens the feedback loop between audience data and the algorithmic (and at times semi-automatic) optimization of that audience data.

2.2. Newsroom Technology Adoption

At an organizational level, resources and culture shape the adoption of technical tools. Since adoption processes impact the effect that technology has on the editorial workflow, they are important for understanding why technology is used in different ways in different newsrooms (Boczkowski, 2004). One driver for adoption is financial, as large quantities of audience data allow for more precise targeting by advertisers (Tandoc, 2014; Thurman, 2011). A lack of financial resources can also dampen the use of analytics technologies. Companies without the means to hire specialists must utilize the tools already available to them. Many journalists who use web analytics in their work learn how to do so informally, and those who enter audience engagement often come from roles that were eliminated or made obsolete (Assmann & Diakopoulos, 2017; Tandoc, 2014). Separate from available resources, organizational priorities and dynamics also determine how tools get used. While some organizations use web analytics as a primary driver in editorial decision making, others limit the use of and exposure to audience data in the newsroom (Anderson, 2011). Those relative prioritizations manifest themselves in tools and in support structures (Nielsen & Cherubini, 2016).

At the receiving end of those pressures are individual journalists, who must go through the negotiation of adopting new technology while maintaining their traditional roles and responsibilities (Tandoc, 2014). Journalists see data as a more objective source of feedback than other methods, one that increases editorial efficiency (MacGregor, 2007). In a 2013 survey of journalists, 90.5% said that reader data "have helped them serve the audience" (Jenner & Tandoc, 2013). A positive attitude is especially prevalent in those who focus primarily on what the audience wants, or who tie increased readership to economic gain (Vu, 2014; Zamith, 2018). Journalists with more traditional conceptions of their gatekeeping roles feel a strain between the value of news judgment and the push to use data in the reporting process (Anderson, 2011).

While engineers have blogged about in-house systems used for multivariate headline testing (Arak & Kentaro, 2017; Muralidhar, 2016), and editors have circulated guidelines for writing effective headlines (Gessler, 2016; Rayson, 2017), neither camp of practitioners addresses the individual and organizational factors this research is concerned with. In most cases, the popular and scholarly literature only mentions headline testing as an existing practice in digital newsrooms, failing to delve into the specific details of how the practice changes headline writing, editorial and ethical thinking, and organizational perception (Christian, 2012; Kuiken, Schuth, Spitters, & Marx, 2017; Reisman, 2016). This research aims to fill this gap by examining how the technology of A/B headline testing is being adopted and integrated into newsroom practices.

3. Methods

We conducted a qualitative study consisting of semi-structured interviews with editorial and strategy staff in newsrooms practicing some form of headline testing, as well as product staff and a third-party vendor that supplied A/B testing tools. We recorded and transcribed each interview and used those transcripts as the basis for a grounded theory approach to data analysis (Glaser & Strauss, 2009, pp. 21–43).

3.1. Participants

After IRB approval was obtained, participants were recruited using three strategies. First, we identified news organizations that use A/B headline testing by searching for blog posts that detailed these processes. In addition, we looked at homepage source code for evidence of sites using headline testing. We then sent recruiting emails to the individuals at those organizations whose LinkedIn profiles indicated that they were most likely to be involved in the headline testing process, in roles such as audience development, homepage production, or analytics. We recruited five participants with this method. Second, we created a survey form that potential participants could fill out with their contact information and shared it on Twitter and to targeted digital media industry groups on Facebook and Slack. We contacted respondents who self-identified as media company or third-party vendor employees. This method brought four more participants. Finally, we utilized snowball sampling, asking each participant for recommendations of further potential interviewees. We recruited one additional participant this way. Opening up recruitment to self-nominations via social media and to snowball sampling allowed us to reach relevant individuals who were outside our initial preconceptions of who might be involved in headline testing within news organizations.

After contacting 32 newsroom workers in relevant roles, we recruited and interviewed 10 participants for this study, representing a mix of perspectives from audience development, social media, editorial, product, and marketing staff. We noticed similar responses around the fifth interview and reached theoretical saturation after the seventh. Three more interviews after this point confirmed our findings. Participants provided a cross-section of the digital media landscape, as representatives of legacy news organizations (N = 3), established digital publishers (N = 2), smaller topical blogs (N = 2), and larger industry- and interest-specific digital outlets (N = 2), as well as a third-party vendor (N = 1). Of the news organizations represented by participants, five had formal testing systems in place. Three newsrooms used informal approaches to headline optimization devised by the organization, and one was in the process of building out headline testing capabilities. Seven of the participants were women, and three were men.

3.2. Materials

Our interview script consisted of 14 questions, with several potential follow-up prompts for each. We developed the script by reviewing existing literature on the adoption of technical tools in newsrooms and journalists' perceptions of their audiences. Topics for each interview included participants' background, general headline writing and testing processes, takeaways from headline A/B tests, the technical tools used in these processes, and ethical concerns around headline testing. Since our participants came from different roles and perspectives within each organization, the semi-structured interview approach allowed us to tailor interviews to each participant's expertise through unscripted follow-up questions.

3.3. Procedure

We conducted audio-only interviews in August and September 2018 in English via Skype. Interviewees received information about participating in the study via email ahead of each call, and we obtained consent verbally. Calls were recorded for transcription purposes. The median interview length was 52.5 minutes (max = 60; min = 22). No monetary incentive was provided to participants.

3.4. Data Analysis

We transcribed each interview immediately after completion. We then used key excerpts from the interviews as the basis for an initial low-level coding of the data. These codes were grouped into larger categories, an iterative process that then informed future interviews. As new data became available after each interview, we used a grounded theory approach to further build out and layer categories, with the ultimate goal of developing core categories and ensuring our analysis was well-integrated with the data (Glaser, 1965).

4. Findings

Two overarching areas of focus emerged from our analysis. The first revolves around headlines themselves, detailing the technical nature of testing tools and how they interact with headline writing best practices. The second looks at the social dynamics surrounding headline testing, which play an important role in determining how effective the process can be in a newsroom. Both highlight the context-dependent nature of headline testing.

4.1. Pursuing Better Headlines

The goal of A/B headline testing in all cases was increased traffic to stories. Participants stressed that they wanted to accomplish this goal by improving the quality of headlines, as judged by how well they communicated the contents of the story and adhered to the publication's style and tone. Markers of quality were communicated by way of established best practices, articulated as components of institutional preference and personal experience in digital media, and informed over time by the results of headline tests. The perceived connection between headlines that conformed to editorial understandings of quality and headlines that drew in more readers created a largely harmonious testing process in participants' newsrooms. However, there were points where considerations of quality and traffic diverged, sometimes leading to tensions around the definition of a "better" headline in those otherwise neat priorities. To the extent that it was technically feasible, newsroom workers also incorporated their understanding of various contexts a headline could appear in, such as social media or search results, into evaluating quality.

4.1.1. Tool and Testing Mechanics

A/B headline testing systems work by showing different portions of a site's audience different headlines for the same story. To run a test, potential headline options (often between two and five) for a story are typically written by a writer or editor and loaded into the testing system. During the test, when a user visits a designated part of the site (usually the homepage), the system serves them one of the headlines. This process often continues until the test reaches statistical significance (i.e., one headline can be confidently declared better with respect to the optimized metric, such as clickthrough rate). The system then reports how headlines performed according to the optimized metric. Some systems automatically begin showing all users the winning headline, while others report data on test results and leave the decision of which headline to display to newsroom workers.

There were no drastic variations in how organizations approached headline testing. While some formal systems were third-party tools and others built in-house, they all provided the same testing functionality. Three participants only had access to informal headline testing or optimization approaches. In these cases, they manually made adjustments to the canonical headline of the story. They focused on stories that were underperforming and monitored performance before and after a change to gauge its effectiveness. Whether or not a formal testing system was present in the newsroom didn't correspond with any specific kind of organization.

4.1.2. Toward Best Practices

Participants talked about a variety of headline writing best practices. These ranged from specific, mechanical rules around headline construction (e.g., including salient quotes and numbers, starting explanatory headlines with "how" or "why," referencing important people and organizations by name, and using relevant SEO terms) to more subjective ideals (e.g., highlighting the smartest angle of a story, conveying a story's importance

and timeliness, maintaining a conversational tone, and matching the publication's style). There were no contradictions in what constituted a best practice between newsrooms, suggesting the emergence of a consistent set of data-driven beliefs about headline writing.

The best practices described above were largely revealed or affirmed by headline tests. Tests often contain the same core idea approached using different formulations, as in an example from the *New York Post*, which tested five headlines, including "Is watching porn harmful to your health?," "You'll never guess how much porn Americans watch," and "This is what porn does to your brain" (NY Post Poetics, 2017). These kinds of tests pit different styles and approaches against each other, over time revealing patterns of continued success that grow to define best practices. In uncovering those best practices, then, the emphasis was on taking a longer-term view. This meant cutting back on headline testing over time, driven by a couple factors. First, there was a concern that frequent A/B testing was optimizing at too granular a level, obscuring findings that could be applied more broadly. To combat this, several participants were starting to exercise more discretion over how often tests were run and for what purpose. For example, one participant had a number of long-term questions (e.g., how straightforward a headline should be, or how long) and only selected tests that could contribute to answering those questions. Second, uncovering the answers to those long-term questions naturally led to a reduced need for testing. Best practices remained stable once they were uncovered, meaning headline testing provided diminishing returns.

Headline testing was also mentioned as a training tool, something that taught writers about what works and kept headline writing front of mind. As one participant elaborated:

> It is kind of about the data but it's equally about training junior writers to have stronger angles and write better headlines before they even start writing a story. Generally what wins for us is what's smart and what has an interesting opinion, and that can be hard to train in a junior writer. So if you're working with people on headline alternates, if you're showing the data about what succeeds or fails, that helps train a whole newsroom over time to get better. (Interview, August 9, 2018)

Participants again framed headline testing as something to advance the goal of more traffic, often eliding any explicit mention of a connection between traffic optimization and a tension with ethical normatic considerations. Eight of the participants expressed an aversion to headlines that constituted clickbait or information gapping, the practice of withholding key information in order to get people to click on a story. They described a shift away from these styles in response to audience backlash and toward a more straightforward approach. But because of this characterization of clickbait as a trend of the past, participants seldom reflected on how it might be incentivized by the traffic-oriented focus of headline testing. Rather, they focused on the importance of editorial oversight in the process as a way of ensuring the quality of headlines. Only one participant discussed this tension:

> Some reporters told me that they intentionally game the system to try to do very clickbaity headlines so that they will win, to get them a boost in feed views on our site….So it's tough combating stuff like that. (Interview, September 7, 2018)

In addition, there was little consideration around the roles headlines play apart from attracting clicks. Headlines can shape perceptions of reality, framing the events of a story along a certain narrative and directing public discourse (Entman, 2007). However, there was no special consideration of the ethics of framing decisions in headline testing. Participants were accustomed to selectively broadcasting information about a story in contexts like Twitter, choosing a headline, image, and descriptive text to frame it from a certain perspective. Headline variants were treated with a similar logic. All variants that got tested were considered valid representations of a story, with no concern for how optimization might then privilege certain frames.

Another concern was how much test results actually reflected the impact of changes being tested (i.e., if a longer headline winning over a shorter one was due to its length). Many variables affect how a story performs in addition to its headline. The vendor participant discussed this concern from a different angle: She saw clients stop using their A/B testing integration over time because they had so many other more pressing priorities and sources of data vying for their attention. She highlighted analytics around audience engagement, membership, and video as other high priorities for newsrooms. There seemed to be a general sense from participants that headlines are important, but little quantitative proof of how relatively important optimizing them is for the business over time, compared to other data-oriented efforts.

4.1.3. Technical Constraints

Participants' default approach was to update headlines across every platform and context of publication once a test was completed and conclusive. This was framed as an assumption that certain characteristics of headlines resonate with audiences regardless of where they appear, and that finding those characteristics by means of testing unlocked a universally-effective approach.

There's another plausible explanation for why newsrooms select one headline framing, however. The technical tools necessary to test and change headlines expediently across many different platforms don't exist. Depending on the content management system (CMS), edi-

tors might only have the option to choose one headline across all contexts, or only add a search- or social-specific headline. With these limitations in place, it makes more sense to think in terms of what makes the best headline for a story generally, rather than needing to test across the different contexts a story would appear in. In addition, testing systems are limited in their ability to reveal differences across contexts. No participant conducted totally distinct testing across platforms, and only one did any native testing on a social media platform (Facebook). They either focused on testing for their most important distribution channel—in some cases search, in others social media—or on the broader qualities that made headlines effective in multiple contexts. They recognized the differences in audience preferences and demographics across different distribution channels, but there were limitations to how much those differences could be reflected in headline selection and testing because of CMS and/or platform constraints.

4.2. Social Dynamics

Headline testing is strongly influenced by the social dynamics of the stakeholders of the testing process at every level. First, the delineation of who controls the testing tool and how they interact with other parties determines how efficiently testing can occur. Second, fruitful interpretation and integration of test results depends on productive communication between the owner of the tool and their editorial collaborators. Finally, communication has a much better chance of being productive if credibility and feedback loops are established between parties. These interdependent factors determine how well a headline testing approach can integrate into a newsroom's workflow and impact how headlines are written.

4.2.1. Roles

There was no consistent title for the person who "owns" the A/B testing tool and process within an organization, the person who is in charge of training, interpreting and sharing test results, and overseeing the mechanics of running tests. People who control those aspects include audience development and engagement staff, editorial strategists, social media coordinators, and growth editors, a cluster that one participant summarized as "optimizer type people." For brevity, we refer to these individuals as "optimizers." These participants' time in journalism ranged from less than a year to a couple decades. Some started in traditional editorial roles, in print or digital, and some began their careers as optimizers. None expressed a desire to move into more writing- or reporting-focused journalist or editor roles. In all cases, newsroom participants owned the testing tool at their organization, or, in the absence of formal systems, were the person pushing for tool adoption and headline best practices.

Tool ownership is an important concept because of how roles are situated in the newsroom. In most cases, optimizers considered themselves highly integrated with their editorial staff collaborators but saw an implicit or explicit barrier between themselves and editorial work. While a couple participants held roles with the word "editor" in the title, the word "journalist" never came up in self-descriptions of any roles. That division allowed groups to claim areas of expertise and become territorial about responsibilities. Tool ownership became a way to shift the power dynamic between editorial staff and optimizers, a change that one newsroom participant without a formal testing system in their current newsroom explicitly detailed as a benefit of A/B testing:

> It is a best practice in this sort of role to be deferential to editors and to writers, because they can be very precious, and you want to make sure that those relationships are really strong. But when they're just straight up wrong, it is really useful to be able to have that data to say like, we know that people click on—I'm using images because it's just really easy—We know that people click on pet puppies more than they do on babies. So this story about puppies and babies really should have a puppy photo, if we can't find one that's puppies and babies. And we can back that up with data, and the data is objective. And we both—even though we're sort of having this negotiation of whose position we're going to go with—we have this objective third party who's providing us information that neither of us is disputing. And so it's sort of changed the power dynamic a little bit—I hate to call it a power dynamic—but it's changed the power dynamic a little bit, because we can point to this third party data that we both agree is valid to say, 'I appreciate that you would like to present your story in this way. However, we know that that particular presentation is less likely to be successful than this particular presentation'. (Interview, August 14, 2018)

In this scenario, the deciding factor was a best practice brought up by the optimizer. However, while the best practice was a known, successful approach to presenting content, the optimizer further justified its use with third-party data to defuse a potential confrontation. The framing of reader data in the neutral language of an "objective third party" also allowed the optimizer to implicitly elevate audience preferences over editorial judgment without directly challenging the journalist's position.

Participants were quick to talk about the headline expertise that reporters and editors built over a career in journalism. However, optimizers also had knowledge of best practices for digital headline writing and a desire to enact those best practices in the newsroom. To do that, they needed a way to have conversations about headline writing with editorial workers on equal footing, not as outsiders trying to encroach on claimed territory. By owning the headline testing tool, optimizers elevated themselves to active participants in the publishing process. They offered something that journalists wanted—

more traffic to their stories—and in return got both explicit and implicit control over the headline writing process: Explicit control to select headline variants for testing, and implicit control to push editorial workers in the direction of best practices by demonstrating success with test results. As a result, no participants reported significant pushback from their editorial collaborators on the adoption or use of headline A/B testing.

4.2.2. Productive Communication

Participants identified productive communication with their editorial stakeholders as key to the testing process. The examples they gave of productive communication contained three common characteristics: proximity, consistency, and accessibility.

Proximity: Optimizers stressed the need to work closely with their editorial counterparts. One participant talked about how much of their work occurred at the individual relationship level, building up trust and credibility person by person, while another talked about how important the work of evangelizing the tool was during the process of adopting A/B testing at their organization. While optimizers used less personal communications channels (e.g., Slack messages and email newsletters), face-to-face communication was a key component of maintaining good relationships with editors. This took the form of formal relationships—such as recurring meetings with specified agendas, collaborations on big stories, and embedding with desks—as well as the casual conversation that came from working side-by-side in the newsroom. Proximity could also be achieved in digital communications channels. Three participants mentioned how helpful it was to have their teams present in the Slack channels that editorial workers used for workshopping headlines, as it gave them the chance to suggest changes and tests in the moment.

Consistency: Optimizers communicated results and best practices on a fixed schedule. Whether it was daily automated results, weekly newsletters, monthly reporting, or a recurring meeting, optimizers set expectations for what kinds of data editors should expect and the schedule they should expect it on. These regular communications, when curated by optimizers, typically highlighted specific examples of successful tests and illustrations of broader best practices that optimizers were monitoring. Automated communications provided a record of the results for each headline test. Rather than waiting for editors to come to them with questions about test results or best practices, optimizers proactively provided assistance. This increased the visibility of optimization work, kept headline writing front-of-mind, and encouraged further discussion.

Accessibility: For the data and insights provided by optimizers to be used effectively, it needed to be accessible and legible to those who were interested in its implications. This was accomplished by sharing results in public Slack channels and opening up data from tools like Chartbeat to anyone with an account. Even in cases where data was restricted, limitations were more a case of only sharing data with people who would be interested, rather than one of gatekeeping sensitive information. Optimizers wanted as many people as possible in the newsroom to have access to their work. This accessibility also had the effect of increasing accountability. When knowledge was circulated widely, the impetus to improve headline writing shifted from those discovering best practices (optimizers) to those putting them into practice (editors).

4.2.3. Influence

A big part of maintaining productive relationships with editorial staff for optimizers, in which their suggestions were valued and implemented, was establishing credibility. As one participant acknowledged, journalists tend to be inherently skeptical people who, while generally "not highly numerate" themselves, interact with powerful figures who can fabricate statistics to advance an agenda. Journalists wanted to know that this optimizer could back up his recommendations with credible expertise.

Optimizers achieved that credibility by the training they went through and gave, and through the expertise they held and asserted. The training process itself didn't generally suggest much credibility: four participants either were informally trained one-on-one or taught themselves how to use the testing tool. However, journalists didn't have insight into that process, and the piece that they did have visibility of—that optimizers had spent more, or any, time learning the headline testing tool and process—helped establish their role as experts. Further bolstering this perception was the fact that optimizers tended to be the ones who taught everyone else how to set up, run, and interpret headline tests.

Once established, credibility asserted itself in communication. Participants explicitly characterized communication between optimizers and editorial staff in terms of feedback loops. Both directions of each of these dynamics were important in giving groups a sense of control in the headline testing process.

Editorial to optimizers: In the testing process, editors and reporters exercised control over certain aspects of headlines. In almost every case, headline variants for tests came from collaboration between reporters and editors, or editors were at least consulted. Editors also frequently initiated headline tests. In addition, while in two newsrooms winning headlines were automatically selected by the testing system, editorial workers often controlled the most important form of feedback: the ultimate decision of whether or not test results were incorporated into an article's headline. One participant elaborated on cases where this editorial intervention might manifest:

> Sometimes our opinions of heds will have changed by the time we are making that decision…we have fields

> for both the article page and the social headline, so sometimes we want to have like a more explicit headline on social, and will keep a more, kind of magazine-y—what I might call vague—headline on the article page. It really goes back to the person who requested it to decide what they want to do with the information…we do not default to, 'and then we will change it everywhere to the winner of the test'. (Interview, August 10, 2018)

In contrast, in one of the newsrooms with an automatic system, readers were shown the winning headline by default after a test ended. While the system sent a notification with results to the person who set up the test, there was no human intervention. Since headline variants were often written by writers or editors in this newsroom, though, there was an understanding that editorial judgment had still been exercised at some point in the testing process. Automatically resolving tests was framed as a way to make headline testing more open to all writers and editors, but it also made enforcing rules around quality and best practices more difficult.

Optimizers to editorial: optimizers often provided editorial feedback in the form of test result data. This could be presented as either a single case communicating the results of one test, or as a longer trend. Though they had varying levels of involvement at every stage of testing, they acted primarily as interpreters of, and advocates for, test results.

The one area that participants consistently mentioned as the exclusive purview of editorial staff in A/B testing was the accuracy of headlines. In any negotiation around what alternatives to use in a headline test, or how to change a headline to optimize it better for the web, editors had veto power if they perceived a proposed headline as factually inaccurate or failing to convey the point of the story. Optimizers differed in their perception of this authority. Some were content to acknowledge editors' expertise and work within their requirements. That attitude came from a mix of respecting editorial experience and authority, and as a matter of expediency. One participant negotiated changes to headlines individually with editors for every story, so surrendering judgment over factual accuracy was a way to speed up conversations. However, another participant also felt that editors were overly cautious when thinking about story presentation:

> I think historically, going back to the print side, newspapers were very profitable for a long time...It was kind of in the business' interest to be conservative with a small 'c' in how they presented the news. And to a certain extent, that got confused with ethics in journalism in my mind, that it was ethical to almost be dull. And to me, being interesting- the continuum between interesting and dull is unrelated to the continuum- or not highly related with the continuum between ethical and unethical. To me they're two different things, but I think they got confused. And so there was no market imperative for people to be highly compelling in how they wrote headlines. (Interview, August 20, 2018)

Participants acknowledged that editorial workers had the authority to step in when they felt a headline wasn't factually accurate, but they differed in how willing they were to push back on that judgment with the backing of data and through careful management of editorial relationships for the sake of traffic. In addition, both editorial workers and optimizers had agency in shaping the broader contours of headline testing. Editors could request or initiate tests on their stories, and optimizers could flag low-performing stories for testing.

5. Discussion

Our study advances two main points brought forward in discussions around actor-network views of the newsroom. First, we extend the concept of a mediating actor in the newsroom between other actors and technological actants (Schmitz Weiss & Domingo, 2010). We've detailed two aspects of A/B headline testing: the content of headlines, and the social dynamics of the testing process. The intersection between these areas hinges on the optimizer. By running headline tests and bringing results to editorial workers, and by incorporating the suggestions and feedback of those editorial workers into testing, the optimizer allows editors to interact with the headline testing process without directly manipulating it. This is a similar role to that played by the production team in Schmitz Weiss and Domingo's (2010) case study of *El Periódico*, that of a "buffer" between journalists and technical tools. However, optimizers also raise the potential for "bridge" actors to attain greater agency. While the production team acted as an interpretive conduit to editorial complaints, optimizers are more transformative in their transmission of testing data and exert more direct influence on their editorial collaborators.

Second, by detailing the impact of A/B testing on the headline writing process, we reinforce the role technological actants play alongside human actors in shaping journalism as a sociotechnical phenomenon (Primo & Zago, 2015). While editorial judgment plays a role in shaping the inputs of an A/B testing system and often contributes to final selection on the output side of the system, the technological tool itself constrains and prioritizes output decisions. In some cases, publishers were willing to distribute optimized content with minimal direct human intervention aside from writing headline variants, but meaningful future work remains to be done to understand the conditions under which technical actants may overcome the types of negotiated editorial control that we predominantly observed. The role of the testing system speaks to Primo and Zago's (2015) conception of algorithms as mediating and transformative, a framing that points to a beneficial relationship be-

tween algorithms and human actors in co-creating news. However, those testing algorithms also extend beyond the purview of newsrooms and their ability to negotiate value tensions and exert editorial control, such as in Facebook's recent rollout of an organic testing tool (Moses, 2018). Specifically, relying on technology platforms for testing tools and data could further infrastructural capture, in which news outlets become editorially conflicted because of their reliance on the data and audience platforms provide, and algorithmic isomorphism, in which the dependence news publishers develop on a platform like Facebook for distribution allow the latter to shape the former as it sees fit (Caplan & boyd, 2018; Nechushtai, 2018). As content optimization incorporates additional external actors and actants, future work might examine how editorial control is then negotiated between internal and external actors and actants.

This study also offers evidence related to the benefits of scale sometimes enjoyed by larger newsrooms (Hindman, 2018). Namely, at an organizational level, companies with more resources can afford to hire more optimizers and pay for more testing infrastructure. More optimization presumably means more traffic, audience, and eventually revenue. At an audience level, sites with more traffic reached statistical significance of tests more quickly, whereas some smaller sites in this study stopped testing entirely because of their inability to get meaningful results. Scale becomes a competitive advantage in a data- and algorithm-driven publishing system; smaller publishers may find it difficult to keep up.

Finally, the findings we have presented on how journalists perceive A/B testing interacting with headline writing serves as the groundwork for further quantitative study of headline style and content over time. Our participants expressed little to no ethical concern about the role popularity (i.e., via click data) played in choosing headlines through A/B testing because they saw the effect of data as bounded by editorial judgment. There is room for further scrutiny of this assumption, however. In particular, we see potential in examining how the headline similarity of outlets with explicit focus on traffic generation and those with other stated editorial values has changed over time, as a measure of how traffic pressure has acted on the latter.

5.1. Limitations

There are several important limitations to acknowledge with the sample used in this study that suggest interesting areas for future work. First, because we didn't include writers or reporters in our sample, we're limited in our ability to make claims about inter-role dynamics from the editorial perspective. Moreover, without the benefit of direct observation (e.g., an ethnographic study), we can only take the interpretations participants presented at face value. For example, while participants noted positive reactions to headline testing from their editorial collaborators, that sentiment may come from an optimistic perspective of advocating for the process. Second, our data represents a snapshot of events, attitudes, and perceptions as they occurred at a specific point in time, which precludes any ability to make longitudinal claims or comparisons. Measuring fine-grained changes in headline testing over time suggests yet another avenue for future study. Finally, because the newsrooms in our sample all utilized some form of A/B headline testing or optimization, we cannot make claims about how they compare to newsrooms that don't test headlines. Investigating those comparisons in future work would help provide a broader understanding of how A/B testing interacts with headline writing.

Acknowledgements

The authors thank our participants for sharing their valuable insight, without which this research would not have been possible. This work is supported by the National Science Foundation award IIS-1717330.

Conflict of Interests

The authors declare no conflict of interests.

References

Anderson, C. (2011). Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism: Theory, Practice & Criticism*, *12*(5), 550–566. doi:10.1177/1464884911402451

Arak, J., & Kentaro, K. (2017, August 30). ABRA: An enterprise framework for experimentation at The Times. Retrieved from https://open.nytimes.com/abra-an-enterprise-framework-for-experimentation-at-the-times-57f8931449cd

Assmann, K., & Diakopoulos, N. (2017). Negotiating change: Audience engagement editors as newsroom intermediaries. *#ISOJ Journal*. Retrieved from http://isoj.org/research/negotiating-change-audience-engagement-editors-as-newsroom-intermediaries

Beam, R. A. (1995). How newspapers use readership research. *Newspaper Research Journal*, *16*(2), 28–38. doi:10.1177/073953299501600204

Blanchett Neheli, N. (2018). News by numbers: The evolution of analytics in journalism. *Digital Journalism*, *6*(8), 1041–1051. doi:10.1080/21670811.2018.1504626

Boczkowski, P. J. (2004). The processes of adopting multimedia and interactivity in three online newsrooms. *Journal of Communication*, *54*(2), 197–213. doi:10.1111/j.1460-2466.2004.tb02624.x

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 1–12. doi:10.1177/2053951718757253

Christian, B. (2012, April 26). The A/B test: Inside the

technology that's changing the rules of business. *Wired*, *20*(5). Retrieved from https://www.wired.com/2012/04/ff-abtesting

Cohen, N. S. (2018). At work in the digital newsroom. *Digital Journalism*. doi:10.1080/21670811.2017.1419821

De Sola Pool, I., & Shulman, I. (1959). Newsmen's fantasies, audiences, and newswriting. *The Public Opinion Quarterly*, *23*(2), 145–158. doi:10.1086/266860

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of Communication*, *57*(1), 163–173. doi:10.1111/j.1460-2466.2006.00336.x

Gessler, K. (2016, November 28). 18 tips for writing engaging headlines + 27 makeovers that saved stories from extinction. Retrieved from https://medium.com/@kurtgessler/18-tips-for-writing-engaging-headlines-27-makeovers-that-saved-stories-from-extinction-55b8e73b84a2

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems*, *12*(4), 436–445. doi:10.2307/798843

Glaser, B. G., & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick: Aldine.

Hindman, M. (2018). *The Internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton, NJ: Princeton University Press.

Jenner, M. M., & Tandoc, E. C., Jr. (2013, November 14). Newsrooms using web metrics to evaluate staff, guide editorial decisions. Retrieved from https://www.rjionline.org/stories/newsrooms-using-web-metrics-to-evaluate-staff-guide-editorial-decisions

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, *5*(10), 1300–1314. doi:10.1080/21670811.2017.1279978

Lee, A. M., Lewis, S. C., & Powers, M. (2014). Audience clicks and news placement: A study of time-lagged influence in online journalism. *Communication Research*, *41*(4), 505–530. doi:10.1177/0093650212467031

Lewis, S. C., & Westlund, O. (2015). Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda. *Digital Journalism*, *3*(1), 19–37. doi:10.1080/21670811.2014.927986

MacGregor, P. (2007). Tracking the online audience: Metric data start a subtle revolution. *Journalism Studies*, *8*(2), 280–298. doi:10.1080/14616700601148879

Moses, L. (2018, August 27). Continuing charm offensive, Facebook creates tool to boost news publishers' reach on the platform. Retrieved from https://digiday.com/media/continuing-charm-offensive-facebook-creates-tool-boost-news-publishers-reach-platform

Muralidhar, N. (2016, February 8). Bandito, a multi-armed bandit tool for content testing. Retrieved from https://developer.washingtonpost.com/pb/blog/post/2016/02/08/bandito-a-multi-armed-bandit-tool-for-content-testing

Nechushtai, E. (2018). Could digital platforms capture the media through infrastructure? *Journalism*, *19*(8), 1043–1058. doi:10.1177/1464884917725163

Nelson, J. L., & Tandoc, E. C., Jr. (2018). Doing "well" or doing "good": What audience analytics reveal about journalism's competing goals. *Journalism Studies*. doi:10.1080/1461670X.2018.1547122

Nelson, J. L., & Webster, J. G. (2016). Audience currencies in the age of big data. *International Journal on Media Management*, *18*(1), 9–24. doi:10.1080/14241277.2016.1166430

Nielsen, R. K., & Cherubini, F. (2016). *Editorial analytics: How news media are developing and using audience data and metrics*. Retrieved from https://reutersinstitute.politics.ox.ac.uk/our-research/editorial-analytics-how-news-media-are-developing-and-using-audience-data-and-metrics

NY Post Poetics [nypostpoetics]. (2017, February 16). Is watching porn harmful to your health? [Tweet]. Retrieved from https://twitter.com/nypostpoetics/status/832348940876005377

Pavlik, J. (2000). The impact of technology on journalism. *Journalism Studies*, *1*(2), 229–237. doi:10.1080/14616700050028226

Petre, C. (2015, May 7). The traffic factories: Metrics at Chartbeat, Gawker Media, and The New York Times. Retrieved from https://www.cjr.org/tow_center_reports/the_traffic_factories_metrics_at_chartbeat_gawker_media_and_the_new_york_times.php

Primo, A., & Zago, G. (2015). Who and what do journalism? An actor-network perspective. *Digital Journalism*, *3*(1), 38–52. doi:10.1080/21670811.2014.927987

Rayson, S. (2017, June 26). We analyzed 100 million headlines. Here's what we learned (new research). Retrieved from https://buzzsumo.com/blog/most-shared-headlines-study

Reisman, D. (2016, May 26). A peek at A/B testing in the wild. Retrieved from https://freedom-to-tinker.com/2016/05/26/a-peek-at-ab-testing-in-the-wild

Schmitz Weiss, A., & Domingo, D. (2010). Innovation processes in online newsrooms as actor-networks and communities of practice. *New Media & Society*, *12*(7), 1156–1171. doi:10.1177/1461444809360400

Tandoc, E. C., II. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, *16*(4), 559–575. doi:10.1177/1461444814530541

Tandoc, E. C., II., & Thomas, R. J. (2015). The ethics of web analytics: Implications of using audience metrics in news construction. *Digital Journalism*, *3*(2), 243–258. doi:10.1080/21670811.2014.909122

Thurman, N. (2011). Making 'The daily me': Technology, economics and habit in the mainstream assimilation of personalized news. *Journalism: Theory, Practice & Criticism*, *12*(4), 395–415. doi:10.1177/1464884910388228

Vu, H. T. (2014). The online audience as gatekeeper: The

influence of reader metrics on news editorial selection. *Journalism: Theory, Practice & Criticism*, *15*(8), 1094–1110. doi:10.1177/1464884913504259

Zamith, R. (2018). Quantified audiences in news production: A synthesis and research agenda. *Digital Journalism*, *6*(4), 418–435. doi:10.1080/21670811.2018.1444999

About the Authors

Nick Hagar is a PhD student in the Media, Technology, and Society program at Northwestern University. He researches how platforms and technological tools shape the production, presentation, and distribution of journalism as part of the Computational Journalism Lab. He comes from a background in digital media and is interested in finding ways to make technology more accessible to local news media.

Nicholas Diakopoulos is an Assistant Professor in Communication Studies and Computer Science (by courtesy) at Northwestern University where he is Director of the Computational Journalism Lab (CJL). His research is in computational and data journalism, including aspects of automation and algorithms in news production as well as algorithmic accountability and transparency in journalism. He is the author of the book *Automating the News: How Algorithms are Rewriting the Media* and is co-editor of the book *Data-Driven Storytelling*.