information

MDPI

Article

Anonymity and Inhibition in Newspaper Comments

Magnus Knustad * and Christer Johansson *

Department of Linguistic, Literary and Aesthetic Studies, University of Bergen, 5007 Bergen, Norway
* Correspondence: Magnus.Knustad@uib.no (M.K.); Christer.Johansson@uib.no (C.J.)

Abstract: Newspaper comment sections allow readers to voice their opinion on a wide range of topics, provide feedback for journalists and editors and may enable public debate. Comment sections have been criticized as a medium for toxic comments. Such behavior in comment sections has been attributed to the effect of anonymity. Several studies have found a relationship between anonymity and toxic comments, based on laboratory conditions or the comparison of comments from different sites or platforms. The current study uses real-world data sampled from *The Washington Post* and *The New York Times*, where anonymous and non-anonymous users comment on the same articles. This sampling strategy decreases the possibility of interfering variables, ensuring that any observed differences between the two groups can be explained by anonymity. A small but significant relationship between anonymity and toxic comments was found, though the effects of both the newspaper and the direction of the comment were stronger. While it is true that non-anonymous commenters write fewer toxic comments, we observed that many of the toxic comments were directed at others than the article or author of the original article. This may indicate a way to restrict toxic comments, while allowing anonymity, by restricting the reference to others, e.g., by enforcing writers to focus on the topic.

Keywords: anonymity; inhibition; disinhibition; incivility; toxic comments

Citation: Knustad, M.; Johansson, C. Anonymity and Inhibition in Newspaper Comments. *Information* **2021**, *12*, 106. https://doi.org/10.3390/info12030106

Academic Editor: Diego Reforgiato Recupero

Received: 19 January 2021
Accepted: 27 February 2021
Published: 3 March 2021

1. Introduction

Comment sections are a common feature of online news sites and have been described as a staple of the online experience [1]. Among their functions are to engage readers, to provide a democratic voice to the audience, to document the popularity of the source, and to provide journalists and editors with direct feedback on their published articles. Such information can obviously be very valuable, and based on the number of news sites that host comment sections, news sources want to have this interaction with their audience.

In the past years there has been a movement against anonymous online content, based on the assumption that anonymity leads to hostility and insults, and allows for cyberbullying [2]. To combat unwanted comments, many news sites have adopted a policy that requires commenters to use their real names when commenting. Many news sites do this by using a Facebook plugin, where users must log in to their Facebook account to comment on news articles. Some sites have gone even further by closing their comment sections and use their Facebook pages as the primary platform for user engagement and commenting [3–6], raising concerns about privacy [7].

Many academics studying comment sections are, understandably, focused on the negative aspects of commenting, specifically toxic disinhibition in comment sections. Toxic disinhibition is defined by Suler [8] as online behavior that is rude, critical, angry, hateful and threatening. Based on this definition, this article will use the term *toxic comments* when referring to the subject matter of this study. This included the literature review, where referenced studies may have used other terms. Other researchers have used terms such as *uncivil* and *impolite* comments [9]. These two terms are used in the coding scheme developed by Papacharissi [10] that was used in the current study.

Several studies have investigated why seemingly normal people behave in a disinhibited way when commenting, and anonymity is often brought up as an explanatory factor for toxic disinhibition in comment sections [9–12]. There is, however, a methodological problem when studying comment sections. To compare anonymous and non-anonymous communication, some researchers have used laboratory settings [12,13]. While experimental research designs can provide important insights, there is always the question of generalizing results to real-world situations. Other researchers have studied the differences between anonymous and non-anonymous comment sections on different news sites [11], which means that the two experimental groups come from different populations. With such sampling strategies one might risk results being affected by other variables than anonymity. Some researchers have even tried to study the effect of anonymity by comparing data from comment sections and other platforms, such as Facebook [9]. Obviously, there are many other variables than anonymity that could affect results when comparing a comment section to a social media platform.

The current study aims to improve the methodology of studying comments sampled from real-world sources. To study the effect of anonymity in real-world comment sections it is best to sample from platforms with both anonymous and non-anonymous commenters, and, thereby, estimate the effect of the platforms. This would amount to repeated measures on the same platforms, such that most other factors would be constant between platforms. Ideally the difference between the two groups would be whether they are anonymous or not. However, we must also account for individual differences since the same individuals may not comment both anonymously and openly, at least not under the same signature. The current study samples data from The Washington Post and The New York Times, two newspapers with a comment section where users can choose to use their real names or pseudonyms. The comments sampled from these newspapers represent anonymous and non-anonymous commenters on the same platforms. The research question for this study is: are anonymous comments more toxic than non-anonymous comments? In the literature review we see that the existing evidence for anonymity is based on experimental studies and studies where data is gathered from different platforms. In addition, there are other explanatory factors for why individuals may exhibit toxic disinhibition in comment sections. The null-hypothesis of this study is that there will not be a significant relationship between anonymity and toxicity. If there is an effect, how large is that effect?

There have been online communities as early as the ARPAnet, a precursor to the internet from 1969 [14]. In 1973, the Community Memory public bulletin board system was set up in Berkeley, and Internet was then viewed as a way to revitalize democracy and stimulate public debate and social change [15]. The World Wide Web in 1991, and the release of the Netscape Navigator in 1994, led to online editions of newspapers. By the year 2001 there were over 3400 online newspapers only in the U.S. [16]. At the same time, paper editions have declined.

Comment sections emerge as one form of participatory or constructive journalism [17]. Newspaper editors view comments as one of the most successful forms of audience interaction [18], and the intention is to continue supporting comment sections on online publications [19]. According to the Pew Research Center [20] about one in four Americans have contributed to comment sections. As many as 84% of newsreaders read comments, and studies have shown that reading comments can significantly affect readers' perception of public opinion, as well as change their personal opinion [21]. A more recent study found that news readers' perceptions of a news story was influenced more by the story itself than by comments made by other readers [22]. The same study found that the civility of comments did not influence readers' perception of the comment, but it did influence perceptions of the commenter and trust in the information. These findings suggest that how readers perceive and react to comments is a complex issue, but that there is an effect.

Comment sections have been criticized for being places of uncivil and impolite behavior. Papacharissi [10] developed a coding scheme for uncivil and impolite behavior in online forums, which Rowe [9] used to investigate the effect of anonymity in comment

sections. Reported number of toxic comments vary from 4 to 22% [23]. The variation can be due to differences in which sections were studied, the definitions of toxic comments, and methodological differences. Different policies have also been shown to affect the number of uncivil comments [24].

Anonymity is defined by Scott [25] as "the condition in which a message source is absent or largely unknown to a message recipient". There are many reasons why someone would want to remain anonymous, according to Hogan [26]. External pressures may cause someone to express themselves anonymously in order to be treated as any individual, such as when female Victorian writers used male pseudonyms out of fear of being dismissed based on their gender. In a more modern example, the fantasy author Joanne Rowling published the Harry Potter books under the name J.K. Rowling because boys tend not to read books written by female authors [27]. She later published under the male pseudonym Robert Galbraith for a presumably different audience. Another reason for anonymity is internal motivations, where an individual has a desire to adopt a different persona. Functional motivations for anonymity are present when practical concerns dictate that a pseudonym is necessary, such as when other people share your name. Situational motivations arise when someone wants to keep different part of their online separate. Finally, there are personal motivations for anonymity, such as the desire to create an escape from everyday life [26], or when acting as a whistle-blower (cf. for example Wikileaks).

The study of anonymity and its effect on behavior has a long tradition in psychology. As early as 1895, Gustave LeBon studied how individuals take on a collective mindset when they are a part of a crowd, which makes them act differently than they do as individuals [28–30]. While not directly related to anonymity, LeBon's pioneering research was an impactful turning point in the history of psychological research, as it laid the groundworks for how socio-psychological processes could explain (unwanted) human behavior.

Over a half century later, researchers performed experiments to investigate the effects of anonymity. The term deindividuation was created to describe a state in which individuals experience a loss of their individual identity due to the anonymity provided by being in a large group [30]. Festinger, Pepitone and Newcomb [31] found that deindividuation caused by not feeling observed by others allowed test subjects to indulge in behavior from which they were usually restrained. The deindividualized test subjects made more negative comments about their parents than the control group, suggesting a relationship between the degree to which someone is identifiable and their willingness to make negative statements. To investigate deindividuation in real-world conditions, Diener, et al. [32] performed a study on Halloween where they observed if trick-or-treaters would steal candy or money when given the opportunity. Children who were not asked about who they were or where they lived, meaning that they remained anonymous, were more likely to steal. It was also found that children in groups were more likely to steal, pointing to both anonymity and crowd mentality as explanations for unwanted behavior. Modern theories of deindividuation, however, show that anonymity does not necessarily lead to antisocial behavior. Instead, anonymity has been found to lead to increased conformity with group norms, which again can lead to antisocial behavior depending on the norms of the group in any given situation [33]. The Social Identity/Deindividuation (SIDE) Model challenges traditional models of deindividuation that focus on the self being the basis of rational action and the group serving to impede the operation of such selfhood [34]. The SIDE-model emphasizes the effect of social identities on deindividuation. Reicher, Spears and Postmes [34] argues that anonymity within a group does not lead to uncontrolled behavior, but instead gives the members of a group the opportunity to "give full voice to their collective identities." This may also be negative, in that a group may more forcefully side against its opponents and inflate the sense of consensus on, and legitimacy of, the opinions within the group.

With computer-mediated communication came new opportunities for anonymity. It could be argued that online anonymity is an important requirement for our online lives. Having multiple identities when communicating with different people, such as friends,

family or coworkers, is part of the human social experience. Throughout history it has been possible to share different identities depending on the social context [35]. According to role theory in social psychology we juggle different social roles, implying that having multiple personalities is a normal part of human nature. This is presumably true online as well [36]. As Hogan [26] points out when describing situational motivations for anonymity, when people use their real names it makes it possible for two completely different posts from different sources to be presented in the same search results. A pseudonym makes it possible to avoid context collapse, a phenomenon described by Marwick and Boyd [37] as an online situation where multiple audience flatten into one, making it impossible to differentiate self-representation strategies. In addition, contributors to forums at online news sites—especially the most frequent contributors—support anonymity, expressing positive views about how anonymity promotes freer and livelier conversation [38].

Anonymity involves not being held accountable for one's actions, which seems to underlie most concerns about anonymity [25,39,40]. Alongside alcohol consumption and social power, anonymity has a disinhibited effect that emerges from a common psychological mechanism; lower activation of the Behavioral Inhibition System [41]. These three factors may combine to escalate the effect of disinhibition. There is also a social factor to anonymity, in that other people being anonymous may lead to a person behaving in a toxic way [36]. Postmes et al. [42] found that anonymous group members are more likely to be affected by social influence, meaning that anonymous internet users are more likely to behave in an uncivil manner if others are uncivil.

Several studies have concluded that there is a relationship between toxic disinhibition and anonymity. Rowe [9] found that there was more incivility in comments on the Washington Post comment section than on the same articles on Facebook. This finding was explained by the fact that users of the Washington Post comment section are anonymous. This explanation, however, disregards other possible differences between the two platforms. While the Washington Post provides a standardized comment section, which allows for little functionality beyond commenting on articles, Facebook is a diverse social media platform where commenters do not even have to access an article to comment on it. In addition, it is not guaranteed that commenters on The Washington Post are anonymous. While the Washington Post allows for anonymous commenters, a commenter may also use his or her real name. Furthermore, when you make a comment on Facebook, all of the people in your friend list may not only potentially see your comment but be algorithmically directed towards it. This may restrict free expression, as writers know that someone who knows them may judge them. Obviously, even non-anonymous comments will tend to be more vapid on such a medium, possibly even ameliorated by a tendency to virtue signaling towards people you know socially. While Rowe's findings are interesting as a study of platforms, it is difficult to make definite conclusions about the effect of anonymity in general from his results. In another study, Rowe explores the deliberative value of comments on Facebook and The Washington Post, and concludes that comments left by website users were more deliberative than those left by Facebook users [43].

Dillon, Neo and Seely [44] found that comments from two news sites using a Facebook plugin were less civil and polite than those found on two news sites where commenters could comment anonymously. While this is an interesting result, it is possible that the results could be affected by the fact that the anonymous and non-anonymous comments were sampled from different sources. As the researchers point out in their discussion, "We did not take socio-democratic factors such as the political climate of geographical regions into consideration when choosing the four newspapers."

Santana [11] found a significant relationship between anonymity and civility when studying comments from three news sites allowing for anonymity and eleven news sites where commenters had to use their real names. Though, it is worth noting that Santana's results are based on studying anonymous and non-anonymous commenters in different populations. In another study, where 4800 comments were sampled from 30 news sites, Santana found that anonymous commenters were more likely to write uncivil comments [45].

While the higher number of sources compared to the three sites used in the 2014 study, the anonymous and non-anonymous comments are still sampled from different sites, and other interfering variables cannot be excluded.

The Huffington Post provides an interesting case study of anonymity. In its early days, the news site allowed users to comment using any chosen name. In December of 2013 the site changed its policy so that users had to authenticate their accounts through Facebook, while still allowing them to use a pseudonym to comment. In June of 2014, the site changed its policy again, this time implementing a Facebook plugin, meaning that users had to use their real names when commenting. In a large-scale study of comments from before and after the first change of policy—before and after they implemented a requirement of identification through Facebook in 2013—Fredheim, Moore and Naughton [46] found that comment quality improved after users had to authenticate their accounts. However, a similar study on the comment sections of Huffington Post [47] complicates the issue, as they found that the quality of commenting, measured by the cognitive complexity of comments, improved after the first change of policy, where users had to authenticate their accounts but could still use pseudonyms. However, the second change, when users had to use their real names when commenting, caused a decrease in the quality of discussions. Interestingly, after both reforms the quality of discussions improved over time. This indicates that the durability over time is a more important factor than whether using a real name or a pseudonym.

Lapidot-Lefler and Barak [12] found in an experimental research design that anonymity influenced the numbers of threats made by research participants. Anonymity was, however, not found to influence self-reported flaming, negative atmosphere or flaming-related expressions. Barlett, Gentile and Chew [48] used a longitudinal design involving questionnaires, and found that the more people feel that they are anonymous the more likely they are to cyberbully others. Zimmerman and Ybarra [13] found in an experimental research design that anonymous participants were more aggressive than those who were not anonymous. However, it is difficult to judge the effect size relative to other factors.

While there is some evidence to suggest that anonymity leads to toxic disinhibition, some studies have not found this relationship, which indicates that the effect size might be relatively small. Bae [49] found in an experimental research design that anonymity led to a greater feeling of in-group similarity and more attitude change, but less flaming and fewer critical comments. This result seems to directly contradict the other studies mentioned above, but one explanation could be in the topic of conversation and the purpose of the communication. Imagine a meeting, where all are anonymous and dealing with a problem in common. Such a meeting may be conducive of empathy even for complete strangers. Thus, it is not unconceivable that anonymity may enhance empathy between individuals, and recognizing others as more self-similar, especially if other attributes, such as social class, are hidden.

Researchers have suggested other possible explanations for toxic disinhibition. Berg [50] studied the effect of *issue controversy* and found that it had a greater impact on discussion quality than anonymity, suggesting that even if anonymity leads to a decrease in civility and politeness, what is debated (the topic) has a greater effect than if the debaters are anonymous or not. These results are supported by Ksiazek [51] who found that more people commented on certain topics, and that certain topics were more likely to result in uncivil discussions.

Suler [8] and Suler [36] suggested several explanations in addition to anonymity. *Invisibility*, the feeling of not being seen by those one communicates with, regardless of one's anonymity, is thought to be one possible factor contributing to disinhibited behavior. Another suggested contributor, *asynchronicity*, removes the constant feedback-loop of face-to-face communication. *Solipsistic introjection*, when a person reading a message experiences it as a voice within his or her head, can make the sender of the message become a character within one's intrapsychic world. *Dissociative imagination*, which refers to when one has the experience of the created character existing in a different world, may result in online

interactions being experienced like a game. *Attenuated status and authority* due to lack of real-world ques of status and authority may also be a factor, especially with regards to how commenters react to moderators. *Perceived privacy* may cause commenters to experience themselves as being in a private encounter online, when they should know better. Finally, *social facilitation*, where the social environment reinforces or fail to counteract disinhibited behavior, is thought to be an important contributing factor to disinhibited behavior.

Suler is not the only researcher to point out the possibility of social influence contributing to negative online behaviors. *Conformity*, defined by Gilovich et al. [30] as the changing of behavior in response to real or imagined, explicit or implicit, pressure from others, is a powerful influencer on behavior, in both positive and negative directions. Participants on online bulletin boards have been found to conform by adopting to both positive and negative information posted by others [52]. These findings are supported by Rösner and Krämer [53], who found that a commenter is more likely to write aggressive comments if peer commenters are aggressive. The frequency of commenting may also be a factor in incivility, as frequent commenters have been found to be less civil and less informal [54]. Although, Coe, Kenski and Rains [55] found the opposite to be true, which indicates that there are other factors at play.

While there are certainly many factors that are thought to be contributing to toxic disinhibition, there is evidence to suggest that anonymity is an important factor. Moreover, anonymity is a popular topic of discussion among researchers, and in the media, when trying to explain toxicity. However, the evidence is inconclusive. Previous studies in experimental settings may not correctly reflect natural conditions. In studies sampling data from online sources, different platforms or populations may influence the results. The current study aims to ameliorate this by sampling comments from similar sources that allow for both anonymous and non-anonymous commenters. However, this is not without problems. We will therefore use random effects to identify sources of variance.

2. Materials and Methods

Two online newspapers were chosen to sample comments from: *The Washington Post* (WP) and *The New York Times* (NYT). These newspapers were chosen because, unlike newspapers that use a Facebook-plugin as a comment section, WP and NYT have comment sections where users must create a separate account. During the account creation they must choose a username, which can either be their real names or a pseudonym. This means that commenters on these platforms make up a population of anonymous and non-anonymous commenters who are all commenting on the same articles on each platform. This reduces the likelihood of interfering variables, such as the affordances of different platforms, with different rules of conduct, moderation and different comment section cultures. Both news sources are east-coast, national, fairly mainstream, left-leaning newspapers [56,57]. Despite apparently using different technologies for moderation, the two newspapers have a similar moderation policy and rules of conduct. Therefore, it is expected that differences in toxic disinhibition can be more stringently and reliably associated with the anonymity of the commenters.

Constructed week sampling was used to create two constructed weeks from February of 2018 to February of 2019 for each newspaper being studied. This involved selecting two random Mondays, two random Tuesdays, etc., during the specified timeframe. This method of sampling is recommended for studying daily newspapers because it creates a randomly selected issue for each day of the week. The events during these days are likely to be referenced in both sources. Two constructed weeks have been found to be sufficient for representing a year's content [58]. In total, 39 articles on politics from the randomly chosen dates were chosen for study. The articles were found using Google's advanced search functions, where one can search for results from a specific website (e.g., nytimes.com), date and subject matter (e.g., politics). There were two requirements for an article to be chosen; (1) the article has to be about politics, and (2) the article must have a substantial number of comments so as to ensure that the data included enough comments from each

article to represent the diversity of comments and commenters found in a given comment section. In total, 2451 comments were collected individually and added to a database built for the purpose of securely storing the research data. There were 700 comments were sampled from each newspaper, or 1400 comments in total. During the collection process each comment was coded as being either anonymous or non-anonymous based on their username. From this pool of data, 50 comments were randomly selected for each day and each newspaper, totaling 100 comments for each of the 14 days in the constructed weeks. This adds up to a total number of 1400 of comments sampled for analysis.

The chosen research method for this study was content analysis, which involves establishing categories and counting the number of instances of each category [59]. In this study there would be only two main categories: toxic and neutral. To determine the toxicity of comments they were coded using a coding scheme developed by Papacharissi [10] and used by Rowe [9] was used to categorize the sampled comments. This coding scheme contains 12 categories of uncivil and impolite comments: *threat to democracy, threat to individual rights, stereotypes, name-calling, aspersions, implying disingenuousness, vulgarity, pejorative speak, hyperbole, non-cooperation, sarcasm* and *other* (see appendix for further detail). In the current research, a comment will be labeled as toxic if it fits into any of the 12 categories. In addition to the categories, the coding scheme includes a dimension referred to as *direction.* There are three directions: (1) *Interpersonal* are those comments directed at another commenter; (2) *Other-directed* are comments directed at a specific person or group not present in the comment section; (3) *Neutral* comments are not directed at any specific person or group.

To ensure reliability when determining if a comment was toxic, two coders categorized all 1400 comments. During the coding process, neither coder knew if a comment had been made by an anonymous or non-anonymous commenter, as this information was not presented to the coders during the coding process. After the coders had categorized the comments individually, inter-coder reliability was calculated using Cohen's Kappa a, which is recommended by Hsu and Field [60]. The coders agreed on 91% of the comments, and the inter-coder reliability was found to be 0.73. After the coders had individually coded each comment and inter-coder reliability had been calculated, the two coders met to discuss those comments that they did not agree upon. During this process, the contested comments were discussed, and the coders came to an agreement of which category they both agree on before the final statistical analysis of the data (the detailed instructions to coders are found in Appendix A).

After coding was completed, the data set was analyzed using two methods. An overall association test based on the chi-square test was performed on the coded data to determine if there is a relationship between anonymity and toxic comments. A general linear mixed effects model was developed that used a binomial distribution and a logistic linking function. The formula for the testing involved a linear regression analysis with a dependent variable *toxicity* (yes/no, 1 or 0) being predicted by independent fixed factors anonymity (yes/no), media (NYT/WP) and level (first level or sublevel). All toxic comments were categorized for the direction of the comment either interpersonal (i.e., other commenters), others (including public figures) or neutral (i.e., directed at no particular entity). The model also used commenter identity (a coded signature for anonymous, a coded name for non-anonymous) and the date the comment was written (which is linked to events that happened that day) as random effects to quantify these sources of variance. Random effects assume an open set, i.e., it is assumed that there are many other commenters and many more dates. Fixed effects assume that we deal with a close set, i.e., we are dealing with either anonymity or non-anonymity, either NYT or WP, and it is either first comment or a later comment. This affects how variance is handled by an algorithmic implementation of a general linear model (cf. lme4/glmer, [61]). The results will be presented as odds-ratios, compared to a baseline.

3. Results

Of the 1400 comments, 1181 were written by anonymous commenters and 219 were written by commenters using a real name. When analyzing at all comments from The Washington Post and The New York Times, we see that of the anonymous commenters, 30.7% (n = 363) wrote comments that were coded as toxic. Of the non-anonymous commenters, 20.5% (n = 45) wrote comments that were coded as toxic. In the first analysis, a statistically significant relationship was found between the two variables anonymity and toxic comments ($\chi^2 = 9.3$, $p < 0.002$). The comparison of the count and expected count of anonymous and non-anonymous toxic comments suggests that this relationship is due to non-anonymous commenters being less likely to misbehave in the studied comment sections. Table 1 shows the number of comments for each condition, as well as the expected count if there was no relation between the variables. Analyzing the Washington Post and the New York Times separately produced a similar result, with non-anonymous toxic comments being underrepresented.

Table 1. The count and expected count of anonymous and non-anonymous comments that were coded as toxic and neutral comments.

	Not Anonymous		**Anonymous**	
	Count	Expected	Count	Expected
Toxic	45 (20.5%)	63.8 (29.1%)	363 (30.7%)	344.2 (29.1%)
Neutral	174 (79.5%)	155.2 (70.9%)	818 (69.3%)	836.8 (70.9%)

Below is an Extended Cohen-Friendly graph (cf. [62,63]) that illustrates associations between A toxicity and B anonymity, assuming that all data points are unique examples of comments, but not accounting for writers and dates as sources of variance, which will be analyzed later. The expected number of comments is shown by the dotted lines. The width of each box represents the number of comments in each condition, and their height represents deviation from expected counts. The figure shows that the number of signed toxic comments is significantly lower than expected, suggesting that differences in toxicity between anonymous and non-anonymous commenters could be explained by non-anonymous commenters being less toxic than expected. Significant cells are marked in red. The intent of using association plots is to motivate a more detailed analysis.

As can be seen in Table 1 and Figure 1, there are slightly more toxic comments written by anonymous commenters than expected. However, this is not statistically significant. Rather, for non-anonymous commenters there are significantly fewer toxic comments than expected by chance. In other words, toxic comments among non-anonymous commenters are underrepresented in the data. This variation is statistically significant, indicating that the relational effect is due to non-anonymous commenters behaving better than expected. However, the effect size of the association is tiny (Cramér $\varphi_c = 0.08$).

In the more advanced model, we are able to look closer at sources of variance and we may, therefore, also estimate not only the effect of anonymity, but also the effect of media platform, direction and level of comment, as well as if there is an interaction between anonymity and the strongest other factor.

First, we will examine the data in more detail. Table 2 tabulates comments into neutral and toxic comments for the two websites, divided up by *original and downstream* comments. A comment is labeled an original comment if it comments directly at the article (first position in a thread) and downstream if it is a later comment on a previous comment (adding, following up, simply staying within the thread). We see that comments in Washington Post generally have a higher proportion of toxic comments. This difference between NYT and WP has a small effect size of $\varphi_c = 0.16$ (signed) and $\varphi_c = 0.15$ (anonymous). However, being downstream does not alter the proportion of toxic comments. This will be investigated further using a statistical model.

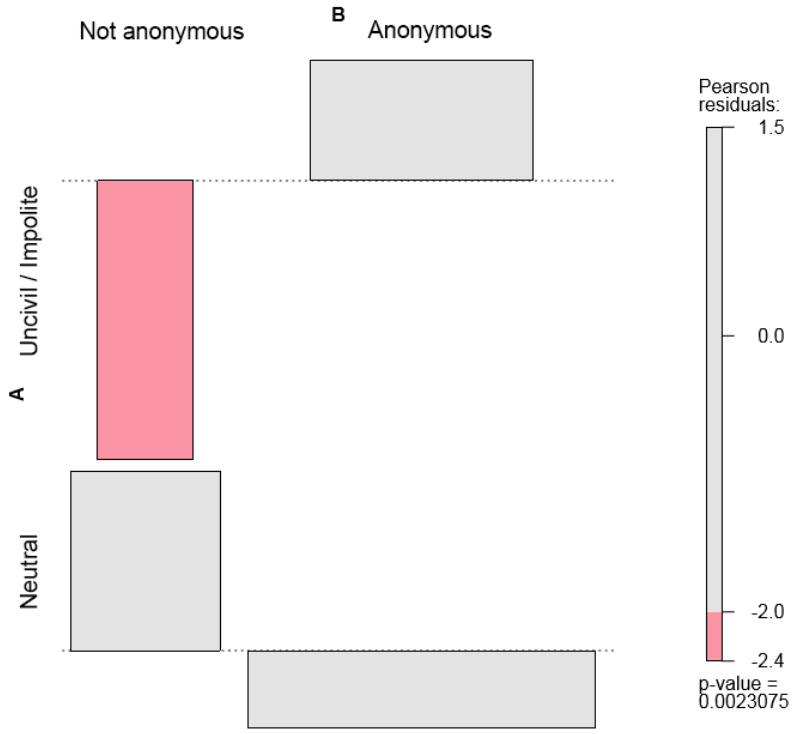

Figure 1. Cohen-Friendly graph of Table 1. Less toxic for signed comments.

Table 2. Proportions of toxic comments in New York Times and Washington Post. For both journals, anonymity is associated with more toxic comments.

	NYT		WP	
	Original	**Downstream**	**Original**	**Downstream**
Signed comments				
neutral	82% (51)	84% (86)	77% (17)	61% (20)
toxic	18% (11)	16% (16)	23% (5)	39% (13)
Anonymous comments				
neutral	74% (142)	79% (271)	54% (143)	69% (262)
toxic	26% (50)	21% (73)	46% (120)	31% (120)

The model is based on a general linear mixed effects model [61] fit by maximum likelihood and using a binomial distribution with a logistic linking function. Commenters may contribute more than one data point, and there are many data points for each date. This will be handled by random effects assigned for identification codes for the commenters and the dates. The Mixed Effect design treats them as sources of variance and may handle these sources simultaneously (more details in Appendix B).

$$\text{Toxic} \sim \text{Anonymity} * \text{Website} + \text{Level} + (1 \mid \text{Date}) + (1 \mid \text{Id}) \quad (1)$$

Formula (1) simply states that we try to explain toxicity in terms of (a) anonymity possibly interacting with website (b) (Response) Level (original/downstream). These are our fixed effects. We are further modeling the sources of variance stemming from (a) the date the comment was written, and (b) the identifier of the commenter. These are our random effects used to control the variance stemming from individuals (id) and events (days).

Caveats: We cannot know if there is only one person behind each identifier. It is possible that more than one person may share a signature, or that an account has been accessed by an unauthorized person. In total we have 1400 data points, and 1083 individuals were identified (Id) in 15 different days (Date). There are relatively few

different dates sampled. However, the dates are considered fairly average dates with no extraordinary events.

Figure 2 gives the odds ratios of our fixed factors. Anonymous is *not* significant ($z = 1.407$ $p = 0.141$) with 45% more toxic comments (1.45). Website is significant ($z = 2.460$ $p = 0.014$) and Washington Post is associated with about 2.61 times the rate of toxic comments. Response level is significant and tend towards *less* toxic comments ($z = -2.357$ $p = 0.018$) for downstream comments. We did not detect a significant interaction between anonymity and website.

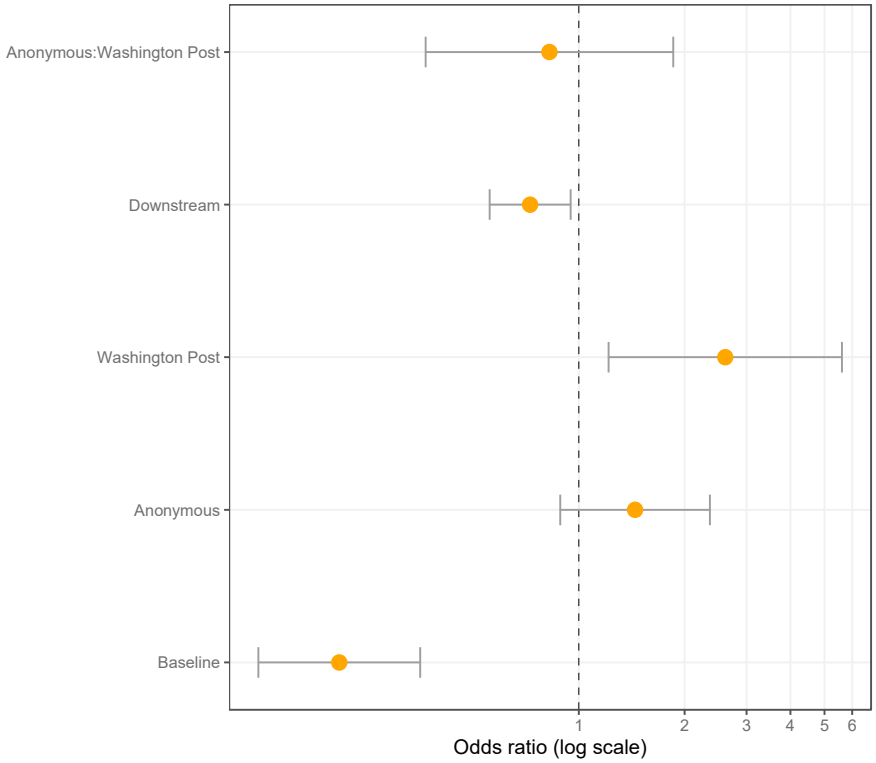

Figure 2. Odds ratio for the fixed factors. Baseline is New York Times, signed, original comments. From Table 3 the baseline is about 18% toxic comments, which is congruent with model estimates ($(20.8 \pm 1.3)\%$). Odds ratio of 1 means no change.

Table 3. Number of toxic comments coded as interpersonal, other-directed and neutral.

Interpersonal			Other-Directed			Neutral		
WP	**NYT**	**Total**	**WP**	**NYT**	**total**	**WP**	**NYT**	**Total**
117	58	175	129	78	207	12	14	26

Table 3 shows that 175 of the toxic comments were interpersonal and directed at other commenters, 207 were other-directed, meaning they were directed at persons or groups not present in the comment section, and only 26 comments were neutral, meaning that they were not directed at any specific person or group.

During the coding process, comments directed at public figures, such as politicians, were specifically marked as being directed at a public figure, in addition to being coded as *other-directed* (Table 4, Figure 3). This subcategory was added to further explore other-directed toxicity. In total, 115 of the 207 other-directed comments were directed specifically towards public figures. While toxic comments directed towards public figures are problematic, there is an argument to be made that the way one speaks about public figures is not the same as when speaking of, for example, other commenters or private individuals. Therefore, we argue that in future research using this coding scheme, the category of

other-directed could be further divided into two categories; comments directed at public figures and comments directed at private individuals.

Table 4. Proportions of toxic comments divided up on what the comments are directed at: 1 is interpersonal 2 is directed at other (including public figures) and 3 is neutral.

	NYT			WP		
	1	**2**	**3**	**1**	**2**	**3**
Signed comments	12% (7)	24% (19)	7% (1)	8% (9)	6% (8)	8% (1)
Anonymous comments	88% (51)	76% (59)	93% (13)	92% (108)	94% (121)	92% (11)

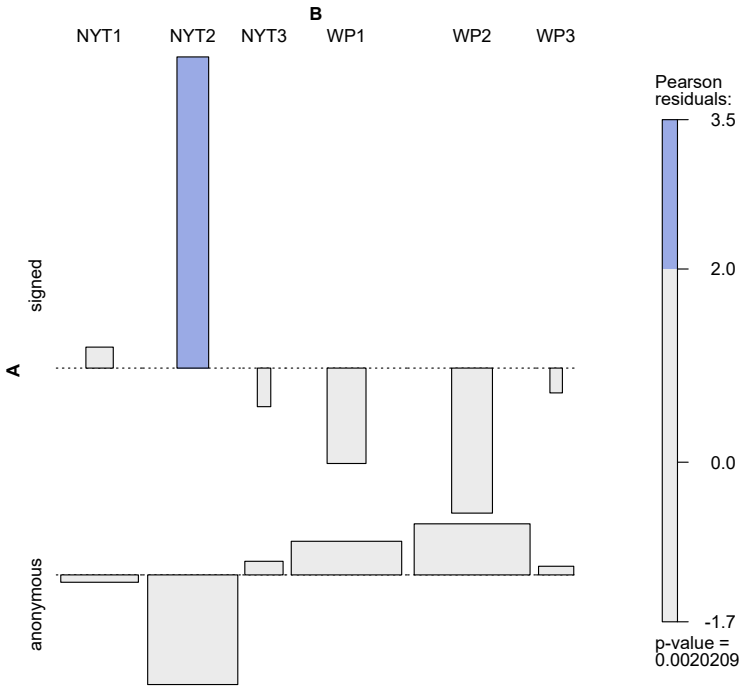

Figure 3. The Cohen-Friendly graph of Table 4 shows that toxic comments directed at others (including public figures) are more associated with signed comments ($p < 0.01$) in the New York Times.

4. Discussion

Comment sections on news sites have the potential to serve as an important channel for public debate. They allow people to express themselves on a variety of topics, with a large potential audience that includes the journalists who wrote the articles being commented on. However, if comment sections are to be a welcoming forum of expression for everyone, it is important to understand why some commenters choose to write toxic and derogatory comments. Our study attempts a contribution to ongoing research on the role of anonymity, but by using a sampling strategy that we believe will provide more accurate results.

We found statistically significant relationships that contribute to understanding toxic disinhibition in the comment sections of The Washington Post and The New York Times. Both can be described as left-leaning media. It is an interesting extension to investigate the effect of political association, on a scale from left to right, but this demands a much larger study. We have decided not to use this dimension, and one motivation is that the political association of the commenters is still unknown.

The result of this study suggests that there is an association between anonymity and toxic comments. Non-anonymous commenters wrote fewer toxic comments than is expected if all were equal. We interpret this to mean that anonymity may have an effect on toxicity, but it is the lack of anonymity that makes a commenter less toxic. In other words, anonymity does not cause toxic comments, but signing a comment either

makes commenters behave better, or possibly signed contributors are associated with more proficient writers.

We found out that there are stronger differences between the two platforms. While anonymity may affect toxicity, editing policies play an equally important role. Because the Washington Post is associated with anonymous toxic comments that website is a stronger explanation for toxicity than anonymity alone. The New York Times may be more active in enforcing their rules of conduct and thus more toxic comments may have been deleted there. The Washington Post and The New York Times have extensive community rules and guidelines that are linked to in the comment sections [64,65]. The rules of conduct themselves give no indication why there would be a difference in toxicity between the two newspapers. While the Washington Post's guidelines are more extensive, both newspapers have guidelines that reflect their desire for civil and well-informed comments, and neither allow personal attacks, vulgarity or off-topic comments. The differences between the two newspapers could be explained by differences in moderation. We do not know how many moderators each newspaper employs, how they work and by what standards they moderate. We do know that the New York Times uses a semi-automated system for effective moderation. In partnership with the Alphabet-owned company Jigsaw, they use machine learning technology for moderation, allowing them to keep comment sections open longer without overextending the resources spent on moderation [66]. It is possible that this system is better at catching unwanted comments than the system used by The Washington Post. The issue of automatic moderation is complicated by the complexity of the task and creative use of language. The state-of-the-art technology for the related task of sentiment detection shows a combined measure of precision and recall between 0.60 and 0.89 (and similar ranges for accuracy) for a wide range of algorithms used on controlled datasets on product and hotel reviews [67]. Chen et al. [68] used Convolutional Neural Networks, with some preprocessing, to detect verbal aggression in Twitter comments with similar results on their test sets. Their test accuracy reached at most about 90% [68]: Figures 7 and 9. Xu et al. [69] show similar results on sentiment detection in comment fields. Algorithms tend to behave worse on truly novel texts outside of the training data, but more data and continuously retraining models may compensate. Even with access to very large databases and deep learning algorithms, there is thus room for either missing a sentiment or mislabeling. In the case of automatic moderation of toxicity, it may create frustration for users if their comments are erroneously publicly flagged or edited out.

It should be noted that the findings in the present study are fairly robust, and the effect of anonymity was detected by different methods. Models that excluded interaction between website and anonymity, and excluded response level, were also tested. The results were very similar. The reason for giving the more elaborate model is to show that other available factors were not responsible for the results. There might, however, be other factors that were not available or controlled in our study.

While the observed relationship between anonymity and toxic comments is interesting, it is important to acknowledge that other associations are stronger, making it difficult to conclude with certainty that anonymity is a significant cause for toxic disinhibition in comment sections. Previous research has concluded that anonymity leads to greater toxicity in comment sections [9,11,44,45]. As mentioned previously, these studies sample data from multiple sources, which could potentially lead to results being skewed by uncontrolled variables. The current study sampled anonymous and non-anonymous comments from the same platforms. While we did find an association between anonymity and toxicity, the result of this study suggests that anonymity has a small effect on the civility of online comment sections. While anonymity may affect toxicity in comment sections, it is certainly not the only factor that should be considered. Issue controversy may play an important role in how commenters debate, as Berg [50] suggested. The comments analyzed for this study were written on political articles at a time of much political controversy and in a highly polarized political climate. However, both anonymous and non-anonymous commenters should be equally affected by political tensions and issue controversy, assuming that people

have honest intentions to discuss the issues. A competing hypothesis is that people choose to be anonymous when they have malicious intent, i.e., intend to disrupt a conversation. This is not supported by our data.

Social influence is another important aspect that should be considered. As stated earlier, conformity has been found to effect toxicity in online communication. It is possible that anonymous and non-anonymous commenters are affected differently by social influence. As noted earlier, anonymity has been found to lead to a greater feeling of in-group similarity and more attitude change [49]. If being anonymous affects a commenter's feelings of similarity to other commenters this could certainly be thought to affect the toxicity of anonymous comments. If we can encourage writers to stay on topic and show more compassion with people or views they do not agree with, then we may ameliorate the negative effects of anonymity, without policing language or opinions.

While the results of this study are interesting, it is important to be aware of its limitations. Firstly, we sampled comments from just two newspapers within a limited time period. Different newspapers use different technological solutions to facilitate commenting, which through affordances, design and moderation policies could be thought to influence the discussions among commenters. Indeed, we detected a significant difference between our two very similar platforms. However, the effect might be platform internal or external. One internal explanation is that platforms, despite having similar rules of conduct, have different editing policies. An external explanation is that the population of commenters may be different between platforms or between levels of anonymity. There may well be larger differences between populations between other platforms, as we chose the examined platforms for their apparently similar political and geographical appeal.

Newspapers use moderators to check for and delete comments that are against the rules of conduct or require deletion for legal reasons. It is possible that comments have been deleted before they could be sampled for this study, and the inclusion of these deleted comments may have had an effect on the results. Therefore, it is accurate to say that our results are limited by an apparent survivor bias.

The comment sections of both The Washington Post and The New York Times allow for users to create any username, and it is possible that some commenters have created pseudonyms that appear to be real names. Obviously fake names, such as *Darth Vader,* were coded as being anonymous during the sampling process. It was not possible for us to verify the identity of commenters using a real-looking name. The websites have more information available; however, sharing such information with a third party violates privacy.

A commenter that wanted to use a pseudonym, would most likely create an obvious pseudonym and not a real-looking name, unless they are sailing under a false flag, which violates standard agreements for setting up a user account. On a platform that allows for pseudonyms, especially one where pseudonyms are the norms, there is little reason for someone to create a name that appears to be a real one.

5. Conclusions

The current study has attempted to improve on the methodology of researching anonymity's effect on toxicity by sampling data from comment sections where anonymous and non-anonymous users debate on the same platform. This novel sampling strategy makes us confident in the results of the statistical tests.

We have found a small but significant relationship between anonymity and toxic comments. At first sight this result seems to support the prevailing view that anonymity causes toxic behavior. However, the data suggest that it is non-anonymous users who are less toxic than expected, and not anonymous users being more toxic. A simpler explanation could be that signed writers are more proficient writers. The effect size of anonymity is tiny or small. Our own analysis showed that the effects of platform and the direction of the comment were stronger than the effect of anonymity. Another interesting finding is the fact that non-anonymous comments were *less* toxic than expected, while anonymous comments were not significantly more toxic than expected. This is congruent with the

observed effect of durable pseudonymity [47], where the quality of comments improved over time for durable pseudonyms. Many anonymous commentors may choose anonymity, not to troll others but to avoid personal attacks in real life. Thus, anonymity is valuable for a freer more democratic debate, and the quality of debate may be improved by fairly simple measures, such as encouraging durable pseudonymity.

Previous research has found other explanations for online toxicity, such as issue controversy [50,51] and social influence [36,52,53]. In our opinion, it is important to evaluate the causes of problematic online behavior. One controllable factor, apart from simply editing out toxic comments (or commenters), is to enforce a discussion to stay on topic and not comment on other users or public figures. As discussed, there are also many positive aspects of anonymity that are at risk if anonymity is cancelled. The small reduction in toxicity may negatively affect the expected quality of comments and limit the diversity of opinions.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; formal analysis, C.J.; data curation, M.K.; writing—original draft preparation, M.K. and C.J.; writing—review and editing, M.K. and C.J.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. Study involve investigating comments made in the public domain. Commentors are not identified by neither name nor signature in this article.

Informed Consent Statement: In accordance with the Norwegian Centre for Research Data, who approved the methodology of this study, consent was waived due to the low impact on research subjects, the public nature of the collected data, and the impracticality of gathering consent from already anonymous subjects.

Data Availability Statement: In accordance with the Norwegian Centre for Research Data, who approved the methodology of this study, data is not available.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Code "1" all comments containing a "threat to democracy": A comment ought to be coded as containing a threat to democracy if it advocates the overthrow of the government (i.e., if it proposes a revolution) or if it advocates an armed struggle in opposition to the government (i.e., if the commenter threatens the use of violence against the government). Examples of such threats include commenters suggesting that government efforts to restrict guns, for example, would lead them to take up arms. For example, one commenter suggested that if the government were to enforce the ban on assault weapons and try and take his gun, "they would soon regret it". Similarly, commenters threatening to start a revolution in response to the government implementing policy would also be coded as a threat to democracy.

Exceptions: Should you believe that the threat is sarcastic, please code for 'sarcasm' (11), not a threat to democracy. "Non-cooperation" (8) should also not be confused with a threat to democracy.

Code "2" all comments containing a "threat to individual rights": A comment ought to be coded as containing a threat to individual rights if it advocates restricting the rights or freedoms of certain members of society or certain individuals. Such examples are common when sensitive or divisive political issues are being discussed because commenters often resort to threatening one another or often advocate restricting the rights of groups or individuals they blame for the event which led the issue to being discussed. For example, following a tragic shooting in which a psychologically disturbed individual is implicated, many people are quick to suggest that the rights of mentally ill citizens be restricted, e.g., "They should all be locked up" would be an example of this. Furthermore, supporters of gun-control often blame those who oppose gun-control, for example, for the widespread use of guns and, by extension, such tragic events. In doing so, they suggest that it is they

who are responsible for such tragedies and, therefore, "they have no right to participate in this debate." Exceptions: Threats to individual rights should not be confused with stereotypes (although they might be closely related if the threat being made assumes that all members of that particular group is the same) or with non-cooperation. Refusing to co-operate is not necessarily the same as refusing others the right to participate in the discussion.

Code "3" all comments containing the use of "stereotypes": A comment ought to be coded as containing a stereotype if it asserts a widely held but fixed and oversimplified image or idea of a particular type of person or thing. This includes associating people with a group using labels, whether those are mild—"liberal", or more offensive—"faggot". The use of stereotypes is common when the topic being discussed is highly partisan.

Stereotyping may also involve making generalized assumptions about the thoughts and behavior of certain groups or individuals based on said stereotypes, for example, suggesting gun-owners/supporters are paranoid, liberals/conservatives are less/more patriotic, or immigrants rely heavily upon social security.

Exceptions: The use of the words liberal or conservative are not always used stereotypically. For example, an administration or an individual may be liberal or conservative in their views, but this type of description is not necessarily stereotypical or derisory.

Note: Stereotypes should also be coded for their direction: those intended to offend others should be coded as antagonistic (e.g., "you liberals are all the same. You want to ban anything you don't like and that doesn't suit you.") or neutral if it was used in articulating an argument but without the intent to offend others (e.g., "the liberal agenda has caused a huge rise in regulations across a number of industries").

Code "4" all comments containing "name-calling": (e.g., gun-nut, idiot, fool, etc.). To be coded as name-calling the words used must be clearly derogatory towards the person it is intended for. Exceptions: Be careful not to include words which may be regarded as a stereotype (e.g., liberal). If name-calling is aimed at a group, or the "name" is often applied to a group of individuals, it may potentially be a stereotypical comment (e.g., anyone who owns a gun is an idiot—this groups all gun-owners together, therefore stereotyping them).

Code "5" all comments containing "aspersions": All comments containing "an attack on the reputation or integrity of someone or something" ought to be coded for aspersion. A comment may be coded as including an aspersion if it contains disparaging or belittling comments aimed at other commenters or their ideas. These ought to include explicit efforts to express dismay at others. For example, a comment which reads: "Teachers don't need to be carrying guns! It's stupid!" may be considered an aspersion. A comment which reads: "sheer idiocy" may also be considered an aspersion. Similarly, a comment which reads: "this is a free country that prohibits slavery. Do you have a problem with that?" may also be coded as an aspersion as its tone implies it is not a genuine question, but an attack on a previous comment/idea. An aspersion may be both explicit or implicit.

Code "6" all comments containing "lying": All comments implying disingenuousness (e.g., liar, dishonest, fraud etc.) of other commenters or public figures ought to be coded as lying Exceptions: If a comment casts doubt on the truthfulness of a previous comment or a public figure this does not constitute the use of synonyms for liar. For example, if a commenter writes "that is not true", they are not implying that the other person is intentionally lying, but rather that they are misinformed.

Code "7" all comments containing vulgarity: All comments containing vulgar language (e.g., crap, shit, any swear-words/cursing, sexual innuendo etc.) ought to be coded as vulgar. Comments containing vulgar abbreviations such as WTF (what the fuck) should also be coded as vulgar.

Code "8" all comments containing "pejorative speak": All comments containing language which disparages the manner in which someone communicates (e.g., blather, crying, moaning, etc.) ought to be coded as pejorative for speech.

Code "9" all comments containing "hyperbole": Comments which contain a massive overstatement (e.g., makes pulling teeth with pliers look easy) ought to be coded as

hyperbole. Be careful not to include words which accurately describe events, particularly given that many of the topics under discussion may be described using words associated with hyperbole (e.g., the Newtown shooting may be described both as a "massacre" and a "heinous" act), although these words are not necessarily used to overemphasize it. Hyperbole might be characterized either as a phrase (e.g., barely a week goes by without a shooting), or the overuse of descriptive words designed to emphasize a point (e.g., "It's not the guns that kill but a ticking time bomb of anger seething in society, giving clues & everyone ignoring him until he kills little babies with an illegal automatic weapon. I don't think it was an accident he killed mommy, the Ph.D. & Principal. He was suicidal & homicidal; very common & wanted notoriety. What better way than to kill babies"). Note: many social issues are discussed using language which may be considered hyperbole, e.g., abortion = murder, gay marriage = abomination, etc. It is up to you as to whether you believe the commenter is making an overstatement or just describes it as such.

Code "10" all comments containing "non-cooperation": The discussion of a situation in terms of a stalemate ought to be coded as non-cooperation. Outright rejection of an idea/policy by a commenter should only count as non-cooperation if it involves excessive use of exclamation marks or capital letters for example. For example, a comment which reads: "I'm 48 years old. I retired after 20 years in the military. I went back to college to be a special education teacher. I WILL NEVER CARRY A FIREARM INTO MY CLASSROOM." Find another solution' may be considered non-cooperation. Similarly, a comment which reads: "I hate guns!! I refuse to send my kids to a school where the teachers are armed!!!!!!!" may be coded as non-cooperation.

Exceptions: A simple rejection of an idea/policy should not be considered non-cooperation. Likewise, suggesting that another commenter has no right to take part in the discussion for whatever reason should be coded as "threat to individual rights" insofar as it threatens their right to free speech, not as non-cooperation. Only a refusal to listen or comply should be coded as non-cooperation.

Code "11" all comments containing "sarcasm": "You'll know it when you see it!!"

Code "12" all comments which may be deemed impolite, but which do not fall into any of the previous categories of impoliteness: This category ought to catch any other type of impoliteness that you think is evident and which does not fit into any other category above. This most commonly includes using capital letters to symbolize shouting and the use of blasphemous language. Even comments you believe are impolite in their tone may be coded as "other" (12).

Exceptions: CAPITAL LETTERS, if used for single words, should be assumed to be signaling emphasis. If a phrase or sentence is written in CAPS, this may be considered shouting.

Direction of Incivility:

All uncivil and impolite comments should be coded for their direction, with the exception of stereotypes which should be coded as antagonistic or neutral. Once the type of incivility has been categorized, the direction then needs to be coded. Comments containing incivility and which are aimed at another commenter in the discussion should be coded as Interpersonal (i). Interpersonal comments include those which are explicitly directed at other commenters (e.g., where the comment includes the name of other commenters) or those which address the comments of others, even without naming them. An example of interpersonal incivility may include: "I can't wait to see you on the battlefield someday Leo [another commenter] because that is what it's gonna boil down to . . . .you believe what you want and you should BUT DO NOT FORCE YOUR BELIEFS ON ME". If the comment contains incivility and is aimed at a specific person or group of people not present, the comment is coded as Other-directed (od). In this case, the "other" often refers to a politician (e.g., Obama), a pressure group (e.g., the NRA), a political party (e.g., Republicans), the media (e.g., the Washington Post) or state institutions (e.g., SCOTUS). If the comment contains incivility but does not refer, or imply reference, to another commenter

or 'other', the comment is coded as Neutral (n). Neutral incivility occurs primarily when the commenter disagrees with the content of the article being commented on. An example of neutral incivility may include: "A Bushmaster in a classroom? WTF!!" The direction of a comment is very much dependent on the coders' understanding of whether or not it refers to other comments in the thread or whether it is a stand-alone comment which is not intended as a response. Thus, it is important to be familiar with the content and language of the article to which the comment refers.

Appendix B

Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]
Family: binomial (logit)
Formula:
Toxic ~ Anonymity * Website + R + (1 | Date) + (1 | Id)
Data: magnus
Control: glmerControl(optimizer = "bobyqa")

AIC	BIC	logLik	Deviance	df.resid
1617.9	1654.6	−802.0	1603.9	1393

Table A1. Scaled residuals.

Min	1Q	Median	3Q	Max
−1.2282	−0.6312	−0.4597	0.9508	2.6189

Table A2. Random effects.

Groups	Name	Variance	Std.Dev.
Id	(Intercept)	0.3259	0.5708
Date	(Intercept)	0.2078	0.4559

Number of obs: 1400, groups: Id, 1083; Date, 15.

Table A3. Fixed effects.

	Estimate	Std. Error	z	Pr(> |z|)
(Intercept)	−1.5694	0.2705	−5.803	6.53×10^{-9} ***
Anonymity=1(anonymous)	0.3685	0.2502	1.473	0.1407 (n.s.)
Website = washingtonpost	0.9595	0.3900	2.460	0.0139 *
Response Level = 2	−0.3193	0.1355	−2.357	0.0184 *
Anonymity = 1: Website = Washington Post	−0.1924	0.4138	−0.465	0.6420 (n.s)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table A4. Correlation of Fixed Effects.

	Intercept	Anonymous	Washington Post	Downstream
Anonymous	−0.730			
Washington Post	−0.496	0.515		
Downstream	−0.273	−0.023	0.002	
Anonymous and WP	0.442	−0.606	−0.931	0.005

References

1. Finley, K. A Brief History of the End of the Comments. Available online: https://www.wired.com/2015/10/brief-history-of-the-demise-of-the-comments-timeline/ (accessed on 2 March 2021).
2. Wallsten, K.; Tarsi, M. Persuasion from Below? *J. Pract.* **2015**, *10*, 1019–1040. [CrossRef]
3. Bilton, R. Why Some Publishers Are Killing Their Comment Sections. *Digiday UK* **2014**, *14*. Available online: https://digiday.com/media/comments-sections/ (accessed on 2 March 2021).
4. Ellis, J. *What Happened after 7 News Sites Got Rid of Reader Comments*; Neiman Lab.: Cambridge, MA, USA, 2015.
5. Ramnefjell, G. Dagbladets Kommentarfelt (1996–2016); Dagbladet.no. 2016. Available online: https://www.dagbladet.no/kultur/dagbladets-kommentarfelt-1996---2016/60160514 (accessed on 2 March 2021).
6. Waatland, E. Nettavisen Stenger Kommentarfeltet Med Umiddelbar Virkning. [The Net-Journal Closes Their Comment Field Effective Immediately]. Available online: https://m24.no/erik-stephansen-gunnar-stavrum-kommentarfelt/nettavisen-stenger-kommentarfeltet-med-umiddelbar-virkning/205456 (accessed on 2 March 2021).
7. Reagle, J.M. *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*; MIT Press: Sabon, NY, USA, 2015.
8. Suler, J. The Online Disinhibition Effect. *Int. J. Appl. Psychoanal. Stud.* **2005**, *2*, 184–188. [CrossRef]
9. Rowe, I. Civility 2.0: A comparative analysis of incivility in online political discussion. *Inf. Commun. Soc.* **2014**, *18*, 121–138. [CrossRef]
10. Papacharissi, Z. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media Soc.* **2004**, *6*, 259–283. [CrossRef]
11. Santana, A.D. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *J. Pract.* **2014**, *8*, 18–33. [CrossRef]
12. Lapidot-Lefler, N.; Barak, A. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Comput. Hum. Behav.* **2012**, *28*, 434–443. [CrossRef]
13. Zimmerman, A.G.; Ybarra, G.J. Online aggression: The influences of anonymity and social modeling. *Psychol. Pop. Media Cult.* **2016**, *5*, 181–193. [CrossRef]
14. Hubler, M.T.; Bell, D.C. Computer-mediated humor and ethos: Exploring threads of constitutive laughter in online communities. *Comput. Compos.* **2003**, *20*, 277–294. [CrossRef]
15. Gonçalves, J. A peaceful pyramid? Hierarchy and anonymity in newspaper comment sections. *Observatorio* **2015**, *9*, 1–13.
16. Li, X. *Internet Newspapers: The Making of a Mainstream Medium*; Routledge: New York, NY, USA, 2010.
17. Løvlie, A.S. Constructive Comments?: Designing an online debate system for the Danish Broadcasting Corporation. *J. Pract.* **2018**, *12*, 781–798. [CrossRef]
18. Singer, J.B.; Paulussen, S.; Hermida, A. *Participatory Journalism: Guarding Open Gates at Online Newspapers*; Wiley-Blackwell: Malden, MA, USA, 2011.
19. Stroud, N.J.; Muddiman, A.; Scacco, J.M. Like, recommend, or respect? Altering political behavior in news comment sections. *New Media Soc.* **2016**, *19*, 1–17. [CrossRef]
20. Artime, M. Angry and Alone: Demographic Characteristics of Those Who Post to Online Comment Sections. *Soc. Sci.* **2016**, *5*, 68. [CrossRef]
21. Toepfl, F.; Piwoni, E. Public Spheres in Interaction: Comment Sections of News Websites as Counterpublic Spaces. *J. Commun.* **2015**, *65*, 465–488. [CrossRef]
22. Graf, J.; Erba, J.; Harn, R.-W. The Role of Civility and Anonymity on Perceptions of Online Comments. *Mass Commun. Soc.* **2017**, *20*, 526–549. [CrossRef]
23. Vergeer, M. Twitter and Political Campaigning. *Sociol. Compass* **2015**, *9*, 745–760. [CrossRef]
24. Ksiazek, T.B. Civil Interactivity: How News Organizations' Commenting Policies Explain Civility and Hostility in User Comments. *J. Broadcast. Electron. Media* **2015**, *59*, 556–573. [CrossRef]
25. Scott, C.R. Benefits and Drawbacks of Anonymous Online Communication: Legal Challenges and Communicative Recommendations. *Free Speech Yearb.* **2012**, *41*, 127–141. [CrossRef]
26. Hogan, B. Pseudonyms and the Rise of the Real-Name Web. In *A Companion to New Media Dynamics*; Hartley, J., Burgess, J., Bruns, A., Eds.; Blackwell publishing Ltd.: Chichester, UK, 2013; pp. 290–308.
27. Savill, R. Harry Potter and the Mystery of J K's Lost Initial. Available online: https://www.telegraph.co.uk/news/uknews/1349288/Harry-Potter-and-the-mystery-of-J-Ks-lost-initial.html (accessed on 2 March 2021).
28. LeBon, G. The crowd: A study of the popular mind. In *Crowd*; T.F. Unwin: London, UK, 1908.
29. Minot, C.S. The Crowd: A Study of the Popular Mind. *Psychol. Rev.* **1897**, *4*, 313–316.
30. Gilovich, T.; Keltner, D.; Chen, S.; Nisbett, R.E. *Social Psychology*; W.W. Norton & Company Ltd.: London, UK, 2016.
31. Festinger, L.; Pepitone, A.; Newcomb, T. Some consequences of de-individuation in a group. *J. Abnorm. Soc. Psychol.* **1952**, *47*, 382–389. [CrossRef]
32. Diener, E.; Fraser, S.C.; Beaman, A.L.; Kelem, R.T. Effects of deindividuation variables on stealing among Halloween trick-or-treaters. *J. Personal. Soc. Psychol.* **1976**, *33*, 178–183. [CrossRef]
33. Felipe, V.; Beria, F.M.; Costa, Â.B.; Koller, S.H. Deindividuation: From Le Bon to the Social Identity Model of Deindividuation Effects. Available online: https://psycnet.apa.org/record/2017-56729-001 (accessed on 2 March 2021).

34. Reicher, S.D.; Spears, R.; Postmes, T. A Social Identity Model of Deindividuation Phenomena. *Eur. Rev. Soc. Psychol.* **1995**, *6*, 161–198. [CrossRef]
35. Kirkpatrick, D. *The Facebook Effect: The Inside Story of the Company That Is Connecting the World*; Simon & Schuster Paperbacks: New York, NY, USA, 2011.
36. Suler, J. *Psychology of the Digital Age: Humans Become Electric*; Cambridge University Press: New York, NY, USA, 2016.
37. Marwick, A.E.; Boyd, D. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media Soc.* **2010**, *13*, 114–133. [CrossRef]
38. Rosenberry, J. Users Support Online Anonymity despite Increasing Negativity. *Newsp. Res. J.* **2011**, *32*, 6–19. [CrossRef]
39. Jacobsen, C.; Fosgaard, T.R.; Pascual-Ezama, D. Why Do We Lie? A Practical Guide to the Dishonesty Literature. *J. Econ. Surv.* **2018**, *32*, 357–387. [CrossRef]
40. Stein, E. Queers anonymous: Lesbians, gay men, free speech, and cyberspace. *Harv. Civ. Rights Civ. Liberties Law Rev.* **2003**, *38*, 159–213. [CrossRef]
41. Hirsh, J.B.; Galinsky, A.D.; Zhong, C.B. Drunk, Powerful, and in the Dark: How General Processes of Disinhibition Produce Both Prosocial and Antisocial Behavior. *Perspect. Psychol. Sci.* **2011**, *6*, 415–427. [CrossRef]
42. Postmes, T.; Spears, R.; Sakhel, K.; de Groot, D. Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior. *Personal. Soc. Psychol. Bull.* **2001**, *27*, 1243–1254. [CrossRef]
43. Rowe, I. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *J. Broadcast. Electron. Media* **2015**, *59*, 539–555. [CrossRef]
44. Dillon, K.P.; Neo, R.L.; Seely, N. Civil keystrokes: Examining anonymity, politeness, and civility in online newspaper forums. In *Internet Research 16*; The 16th Annual Meeting of the Association of Internet Researchers: Pheonix, AZ, USA, 2015.
45. Santana, A.D. Toward quality discourse: Measuring the effect of user identity in commenting forums. *Newsp. Res. J.* **2019**, *40*, 467–486. [CrossRef]
46. Fredheim, R.; Moore, A.; Naughton, J. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. Available online: https://dl.acm.org/doi/abs/10.1145/2786451.2786459 (accessed on 2 March 2021).
47. Moore, A.J.; Fredheim, R.; Wyss, D.; Beste, S. Deliberation and Identity Rules: The Effect of Anonymity, Pseudonyms and Real-Name Requirements on the Cognitive Complexity of Online News Comments. *Political Stud.* **2021**, *69*, 45–65. [CrossRef]
48. Barlett, C.P.; Gentile, D.A.; Chew, C. Predicting cyberbullying from anonymity. *Psychol. Pop. Media Cult.* **2016**, *5*, 171–180. [CrossRef]
49. Bae, M. The effects of anonymity on computer-mediated communication: The case of independent versus interdependent self-construal influence. *Comput. Hum. Behav.* **2016**, *55*, 300–309. [CrossRef]
50. Berg, J. The impact of anonymity and issue controversiality on the quality of online discussion. *J. Inf. Technol. Politics* **2016**, *13*, 37–51. [CrossRef]
51. Ksiazek, T.B. Commenting on the News: Explaining the degree and quality of user comments on news websites. *J. Stud.* **2018**, *19*, 650–673. [CrossRef]
52. Cheng, S.L.; Lin, W.-H.; Phoa, F.K.H.; Hwang, J.-S.; Liu, W.-C. Analysing the Unequal Effects of Positive and Negative Information on the Behavior of Users of a Taiwanese On-Line Bulletin Board. *PLoS ONE* **2015**, *10*, e0137842.
53. Rösner, L.; Krämer, N.C. Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Soc. Media Soc.* **2016**, *16*, 1–13. [CrossRef]
54. Blom, R.; Carpenter, S.; Bowe, B.J.; Lange, R. Frequent Contributors Within U.S. Newspaper Comment Forums: An Examination of Their Civility and Information Value. *Am. Behav. Sci.* **2014**, *58*, 1314–1328. [CrossRef]
55. Coe, K.; Kenski, K.; Rains, S.A. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *J. Commun.* **2014**, *64*, 658–679. [CrossRef]
56. Media Bias Ratings. 2019. Available online: https://www.allsides.com/media-bias/media-bias-ratings?field_featured_bias_rating_value=All&field_news_source_type_tid[1]=1&field_news_source_type_tid[2]=2&field_news_source_type_tid[3]=3 (accessed on 18 November 2019).
57. Where Do News Sources fall On the Political Bias Spectrum? 20 December 2018. Available online: https://guides.lib.umich.edu/c.php?g=637508&p=4462444 (accessed on 18 November 2019).
58. Riffe, D.; Lacy, S.; Fico, F. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*; Routledge: New York, NY, USA, 2014.
59. Silverman, D. *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction*; Cromwell Press: Townbridge, UK, 2001.
60. Hsu, L.M.; Field, R. Interrater Agreement Measures: Comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α. *Underst. Stat.* **2010**, *2*, 204–219. [CrossRef]
61. Bates, D.; Machier, M.; Bolker, B.; Walker, S. Fitting Linear Mixed Effects Models using Ime4. *J. Stat. Softw.* **2015**, *67*, 1. [CrossRef]
62. Cohen, A. On the graphical display of the significant components in a two-way contingency table. *Commun. Stat. Theory Methods* **1980**, *A9*, 1025–1041. [CrossRef]
63. Meyer, D.; Zeileis, A.; Hornik, K. Visualizing independence using extended association plots. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 20–22 March 2003.
64. Amenabar, T. Community Rules. 11 June 2018. Available online: https://www.washingtonpost.com/news/ask-the-post/wp/2018/06/11/community-rules/ (accessed on 7 February 2020).

65. Comments. 2020. Available online: https://help.nytimes.com/hc/en-us/articles/115014792387-Comments (accessed on 7 February 2020).
66. Etim, B. The Times Sharply Increases Articles Open for Comments, Using Google's Technology. 13 June 2017. Available online: https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html (accessed on 7 February 2020).
67. Dashtipour, K.; Gogate, M.; Li, J.; Jiang, F.; Kong, B.; Hussain, A. A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks. *Neurocomputing* **2020**, *380*, 1–10. [CrossRef]
68. Chen, J.; Yan, S.; Wong, K.-C. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Comput. Appl.* **2018**, *32*, 1–10. [CrossRef]
69. Xu, G.; Meng, Y.; Qiu, X.; Yu, Z.; Wu, X. Sentiment analysis of comment texts based on BiLSTM. *IEEE Access* **2019**, *7*, 51522–51532. [CrossRef]