Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept

A pre-print of the article that appeared in Correlating self-report and trace data measures of incivility: A proof of concept. Social Science Computer Review, Special Issue "Integrating Survey Data and Digital Trace Data."
https://doi.org/10.1177/0894439318814241

Toby Hopp, Ph.D*
College of Media, Communication and Information
University of Colorado Boulder

Chris J. Vargo, Ph.D
College of Media, Communication and Information
University of Colorado Boulder

Lucas Dixon, Ph.D
Google Jigsaw

Nithum Thain, Ph.D
Google Jigsaw

Abstract: This study correlated self-report and trace data measures of political incivility. Specifically, we asked respondents to provide estimates of the degree to which they engage in uncivil political communication online. These estimates were then compared to computational measures of uncivil social media discussion behavior. The results indicated that those that self-disclose uncivil online behavior also tend to generate content on social media that is uncivil as identified by Google's Perspective Application Programming Interface. Taken as a whole, this work suggests that combining self-report and behavioral trace data may be a fruitful means of developing multi-method measures of complex communication behaviors.

Keywords: incivility, political discussion, toxicity, survey, computational social sciences

Manuscript word count (including body, citations, endnotes, tables and figures): 7,903

There exists broad societal concern that Americans are talking about political issues in an increasingly uncivil manner. Unsurprisingly, scholars have increasingly sought to understand both the causes and effects of uncivil discourse, particularly the relationship between computer-mediated communication and the ongoing enactment of democracy. However, as pointed out by Muddiman (2017), the incivility literature currently features "multiple and often contradictory conceptualizations across projects" (p. 3183). Although recent work has sought to offer an operationally clarified definition of incivility (e.g., Muddiman, 2017), there persist important knowledge gaps as they pertain to the measurement of incivility in online political communication.

In light of the foregoing, this study set out to assess the degree to which self-report measures of uncivil communication habits correspond to actual online user behavior. Such an exploratory effort contributes to the literature in at least two ways. First, prior research on incivility has generally focused on either individual-level perceptions of uncivil communication or the degree to which deliberative spaces feature uncivil or otherwise toxic language. Missing from the literature is an assessment of the ability of online discussion participants to self-identify the degree to which they communicate in an uncivil manner. Such inquiry is important. According to Papacharissi (2004), incivility is an intentional communication strategy that discards "the collective traditions of democracy" (p. 267) in favor of *deliberate* disrespect. By empirically assessing the degree to which self-reports of uncivil behavior correspond to behavioral-level outcomes, this work sought to arrive at a more comprehensive theoretical understanding of political incivility. Second, research in the communication sciences has increasingly adopted computational methods. While such methods allow for novel inquiry into social relations, there exists a paucity of work focusing on measurement issues, particularly as

they pertain to the development of algorithms that can validly and accurately capture individual-level psychobehavioral phenomena. By combining self-reported and behavioral data, we attempted in this study to sketch logical framework that can be used for the development of trace-data indicators that act as valid proxies for variables that have traditionally been captured via psychometric means. In its mature state, we believe that the here-articulated approach could be useful to researchers. Such utility might be most apparent in instances of negative communication behaviors (e.g., incivility), where self-report measures can be polluted by various forms of reporting bias and/or error.

Literature Review

Incivility in Online Political Communication

In the context of political communication, civility is generally defined as a commitment to a set of social norms that mandate respectful communication between two parties, even if those parties disagree (e.g., Papacharissi, 2004). As noted by Coe, Kenski, and Rains (2014), "commitment to civil discourse—the free and respectful exchange of ideas—has been viewed as a democratic ideal from the ancient Athenian forums to the mediated political debates of modern times" (p.658). It should, then, be of little surprise that research suggests that the abandonment of good faith discussion in favor of uncivil, toxic, or otherwise bad faith discussion has a deleterious effect on the discursive processes that are foundational to democracy (e.g., Brooks & Geer, 2007; Chen & Lu, 2017).

As Americans continue to integrate their offline and online lives, there is concern that digital discussion spaces are facilitating increased amounts of anti-democratic communication, and that this communication has the functional effect of harming the civic beliefs and practices that sustain democracy. As illustrated by Rowe (2015), "many sceptics [of the democratic

potential of online discussion] believe that the relatively high level of anonymity that this medium affords users exacerbates disinhibited communicative behaviour, leading to an increase in impolite and uncivil political discussion" (p. 122). Therein, there is reason to believe that online incivility is not constrained to the specific digital spaces where it occurs. Instead, given the structures of reproducibility that govern the Internet, "moments of incivility now spread more rapidly and widely than ever before" (Coe, Kenski, & Rains, 2014, p.658), negatively affecting not only the quality of computer-mediated communication (e.g., Ng & Detenber, 2005), but also the mental models and democratic habits that govern offline social and political interactions (e.g., Mutz, 2015).

Operationalizing Uncivil Online Political Talk

Although there does not currently exist a settled-upon operational definition of incivility, there are reasons to suspect that it is both something more than spontaneous impoliteness (Papacharissi, 2004) and multi-faceted in nature (Coe, Kenski, & Rains, 2014; Gervais, 2015; Santana, 2014). Santana (2014) posited that incivility is reflected in the existence of one or more of the following components: name-calling; threats, vulgarity, foul language, xenophobia, hateful language, bigoted language, disparaging comments on the basis of ethnicity, and use of stereotypes. Sobieraj and Berry (2011) focused their construction of incivility on so-called outrage speech, including factors such as name calling, insulting language, misrepresentation, mockery, emotional language, and ideologically extreme language. Similarly, Coe, Kenski, and Rains (2014) operationalized incivility as existing in five distinct forms: name-calling, aspersion, lying, vulgarity, and pejorative for speech. Vargo and Hopp (2017) measured Twitter-based political incivility as language that was extreme, vulgar, abusive, or otherwise hurtful in nature. Gervais (2015a, 2015b) recently used the above instances of uncivil speech to create four broad

categories of incivility relevant to online communication: (1) invectives and ridicule, (2) hyperbole and distortion, (3) histrionics and obscenity, and (4) conspiracy theory.

Based on the foregoing literature, the current study conceptualized online discussion-based incivility in terms of the following observable behaviors: *use of profane language*, *use of name-calling*, *use of threat,* and *invocation of negative stereotypes*. Profanity is defined as language that is untargeted and is vulgar or crude. This dimension corresponds with the vulgarity/foul language dimensions identified by Coe, Kenski, and Rains (2014), Santana (2014), and Gervais (2015b). Name-calling corresponds with Papacharissi's (2004) definition of incivility as intentional disrespect and addresses the operationalization approaches used in Santana (2014; name-calling and abusive language), Coe, Kenski, and Rains (2014; name-calling), and Gervais (2015a, 2015b; invectives and ridicule). The threat component refers to interpersonal threats made by the communicator and addresses the threat included in Santana's (2014) operationalization of online incivility. Both Santana (2014) and Coe, Kenski, and Rains' (2014) operational definitions of incivility include factors addressing the use of negative stereotypes, disparagement on the basis of race and ethnicity, and the use of racist/bigoted language. As such, the proposed dimension (here, labeled as invocation of stereotypes) covers the use of negative group-based generalizations to categorize/label others.

The Current Study

As illustrated above, online political incivility is an important object of study. With some recent exceptions (e.g., Muddiman, 2017), however, there exists a paucity of research focusing specifically on the measurement of online incivility, resulting in lingering questions pertaining to the construct's theoretical properties. To that end, this study sought, in part, to offer some clarity on the subject by assessing the degree to which self-reported uncivil communication behaviors

are apparent in social media users' observed online behavior by addressing the following research question:

RQ1: To what degree are self-reported behaviors of online political incivility associated with observed uncivil communication behaviors on social media?

Method

Study recruitment was accomplished using Qualtrics. Four sample controls were enforced: an approximate 50/50 gender split; a requirement that respondents be active users of Facebook and Twitter; a requirement that respondents be current US citizens; and a requirement that respondents be 18 years or older. We also requested that participants talk about political/social issues online at least monthly; however, due to methodological limitations, this was not enforced at the point of data collection.[1] Before participating in the study, respondents were told that the purpose of the study was to learn more about how often they shared political news and information on social media. Participants were also provided with a consent form that articulated the parameters of data collection, including the following statement:

> **At the start of the survey you will be asked to link your Facebook and Twitter account to our survey application. This application will be used to gather posts (e.g., wall posts and tweets) that you have made on the respective services. These messages will be collected anonymously, and at no time will the researchers know your identity, or the identities of your friends.**

After agreeing to the study's terms, respondents were asked to authorize a custom application that was used to harvest their Facebook and Twitter data. The application confirmed that respondents were active users of both platforms. If they did not meet the study's criteria, participation was discontinued and any collected trace data was discarded. If respondents were active users of both platforms, their data were retained and they were piped into the survey environment. Self-report and social media data were joined using an anonymous identification code that was assigned by the web application. For Facebook, we did not capture newsfeed

information or friend information. For Twitter, only tweets were collected. The application conformed with both Facebook and Twitter's terms of service at the time of study execution. The project was approved by the University of Colorado's Institutional Review Board on December 21, 2016.

Measures

Self-reported incivility measure. Using the dimensions identified above, a four-item measure of incivility was constructed. All items were on seven-point semantic differential scales where 1=*never* and 7=*very frequently*. To the best of our knowledge, researchers have not previously generated/employed self-report measures of uncivil online talk. The measures employed in this study were therefore developed for the current project. All items were attached to the following introductory stem: *When communicating with others on the Internet, how often do you….* Use of profane language was assessed using the prompt "*use language that is vulgar.*" Name-calling was measured by asking respondents to how often they "*use words or terms to describe others that you would not want to be used to describe you.*" Threat was measured by asking respondents to indicate the frequency with which they "*verbally threaten others*" Finally, invocation of negative stereotypes was measured by asking respondents to estimate how often they "*use negative stereotypes to describe others."*

Because we were interested in exploring the incivility items both individually and as indicators of an overall composite measure, we averaged these items to form a single index of incivility (Facebook analytic sample: M=2.09, SD=1.10, α=.77; Twitter analytic sample: M=2.13, SD=1.15, α=.77; see below for details on analytic sample construction).

Trace incivility measure

Coding for political talk. Because the instant treatment of incivility pertained specifically to political discussion, it was necessary to first screen for political posts. In all, the dataset contained over 1.6 million messages. A supervised machine learning approach was used to identify political content. First, 100 messages were randomly chosen and annotated by two coders on whether they mentioned political talk (0=no, 1=yes).[2] Of the 100 decisions, the two coders disagreed once (pairwise agreement=99.0%; Cohen's Kappa=.80). Next, a random sample of 1,000 additional posts was then selected and distributed to the two coders. Both coders examined the posts to see if they contained political talk. The annotations were used to build a machine learning algorithm inside of the *DataRobot* platform. Of the available algorithms, the AVG Blender, a neural network ensemble model, had the highest performance scores and was chosen.[3] After initial model construction, subsequent rounds of messages were randomly chosen, stratifying across prediction values, to help reinforce learning across both classes. In all, 5,006 annotations were made by the two researchers. Performance metrics were subject to a 10-fold cross validation, each time training on a randomly selected 64.0% (n=3,136) of the data. The final model for political talk had F_1 and AUC scores of .88 and .98 (respectively), suggesting that precision and recall for the algorithm were excellent (Fawcett, 2004). An accuracy rate of 94.7%, a false positive score of 3.7%, and a Matthews Correlation Coefficient of .85 all indicated that the algorithm distributed its misclassifications evenly and was not prone to a specific type of error (Silva, Anunciação, & Lotz, 2011). In all, 9,838 political Facebook posts and 6,679 political Twitter posts were identified.

Annotation of incivility behaviors. To code messages for the presence of subdomains found in the incivility literature, Google's *Perspective* Application Programming Interface (API) was used. The API serves deep neural network-based machine learning models developed with

Google's TensorFlow, a set of libraries that support the development of advanced machine learning algorithms. The Perspective API is comprised of a series of algorithms designed to identify individual behaviors conceptually linked to toxic online discussion. The algorithms have been tested across multiple domains, including the comments section of *The New York Times* and *Wikipedia's* "Talk Pages."[4]

Use of the API placed obvious parameters on our operationalization of incivility as it limited our analyses to behavioral (rather than intentional) aspects of incivility. However, using the available algorithms, we were able to identify four attributes that broadly corresponded to the incivility dimensions identified in the literature. The profane language dimension of incivility was assessed using the API component that targeted *obscene language*. The disrespect/name-calling dimension of incivility was captured using the API component that evaluated content for the presence of *insulting language*. The use of threatening language was coded using the API attribute that assessed the use of *threat*. Invocation of negative stereotypes was assessed using the API component that evaluated content for the presence of *hateful language.* Table 1 provides a brief description of each API component used in this study.

[TABLE 1]

Next, the Perspective API was used to assign a probability value ranging from 0 (very likely to be civil) to 1 (very likely to be uncivil) for each user post across all four incivility features. To assess the accuracy of this approach, we manually assessed a subsample of messages. Any post with a probability score on any of the four incivility attributes >.50 was classified as uncivil. Two human coders then evaluated a randomly selected sample of 600 messages (approximately 3.5% of the corpus). This sample contained 300 randomly selected messages from Facebook and 300 randomly selected messages from Twitter. The initial pairwise

agreement between human coders for the entire subsample was 98.5% (Gwet's AC1 = .98, $95\%_{CI}$=.97,.99).[5] Disagreement between the coders was solved via discussion. Across the sample, pairwise agreement between the human-coded data and the computationally-derived annotations was 95.5% (Gewt's AC1 = .95, $95\%_{CI}$=.93,.97). For Facebook, agreement between the human-coded and machine-coded data was 94.3% (Gwet's AC1 = .93, $95\%_{CI}$=.90,.97). For Twitter, pairwise agreement was 96.7% (Gwet's AC1 = .96, $95\%_{CI}$=.94,.99]). Looking at the 27 cases in which there was human-computer disagreement, 18 instances were API positive/human negative and 9 instances were API negative/human positive.

Having generally determined the absence of systematic or widespread error in the Perspective API's assessment of user posts, we subsequently averaged the probability scores for the posts at the user level, resulting in a raw measure describing the average probability that a given political post created by a given user contained the incivility attribute of interest.

Final trace measure construction. In the collected social media data, users were active for different periods of time. Even among those with comparable site membership durations, users posted dramatically different amounts of content. Because the annotation procedure described above simply returned the average probability that any given political post emanating from a given user contained the incivility attribute under consideration and the fact that the theoretical properties of incivility suggest that it is a purposeful and perhaps habitual mode of political discussion (Papacharissi, 2004), we deemed it necessary to also account for the frequency with which users employed uncivil language. To accommodate this goal, we generated a weight (w) variable to adjust each raw indicator score for frequency of political communication. This weight variable was calculated as follows:

$$\boldsymbol{w = \log(1+p)}$$

where *p* represents the average number of posts created by each user *per year* of user platform activity. Notably, *p* was logged because the dataset contained a number of extreme outliers. To avoid logging a number below 1, a constant value of 1 was first added to the average number of posts generated per year of site activity. The weight variable was subsequently used as a multiplicative term, resulting in the following:

$$\boldsymbol{IA_k = a^{*}w}$$

where IA_k is the final computed indicator-level measure for each of the four (*k*) trace incivility attributes, *a* is the averaged probability that a random political post created by the user contains the incivility attribute of interest, and *w* is the weight variable. To demonstrate the effect of the *w* variable, Table 2 shows (as an example) the final computed values for the Facebook-derived trace profane language indicator. Specifically, in this table, we used the observed sample minimum, mean, and maximum values for both the raw measure of profane language and the *w* variable. As shown, *w* functions in such a way that it increases the final computed value for Facebook-based obscene language in an ordinal manner. In other words, the final computed score for the indicator increases as a function of *w* value strength. Conceptually, this results in a measure where those who frequently posted political content that contained uncivil sentiment received higher scores than either those who posted frequent civil content or infrequent uncivil content. This approach was deemed especially important because the self-report measures of incivility were anchored on scales where 1=never and 7=very frequently.[6]

[TABLE 2]

As was the case for the self-report indicators, the four individual measures of trace-data incivility were also collapsed into a single averaged composite measure (Facebook analytic

sample: M=0.15, SD=0.15, α=.69; Twitter analytic sample: M=0.15, SD=0.17, α=.86; see below for details on analytic sample construction).

Analytic Samples

Not all respondents in the primary dataset generated political posts. In other cases, participants generated political content on one platform, but not on the other. As such, we constructed two analytic samples: one for those that created at least a single political post on Facebook and one for those that created at least a single political post on Twitter. Only those who provided complete survey data on the self-report incivility indicators were included in the analytic samples.[7] Descriptive statistics for the incivility variables for the samples are shown in Table 3. Table 3 also summarizes the demographic makeup of each sample. Figures 1 and 2 show density plots for the incivility measures for both the Facebook and Twitter analytic samples.

[TABLE 3]
[FIGURE 1]
[FIGURE 2]

Analytic Plan

As a first step, we examined the degree to which the self-report and trace measures corresponded to one another in a binary context. For both the self-report and trace measures, we classified any user with a score of more than 1 standard deviation above the indicator mean as uncivil. Second, we evaluated the indicator-level associations by generating a correlation matrix using Kendall's τ, a nonparametric statistic that does not assume multivariate normality. Because Pearson's *r* is a well-known estimate of association, we calculated the *r* equivalent using the equation provided by Kendall (1970; $r = \sin[0.5 * \tau * \pi]$). Using the τ matrix, we also generated a correlation network graph (Epskamp, Borsboom, & Fried, 2018). Correlation networks are visual representations of correlation matrices that draw upon the human eye's ability to visually process

complex information. Finally, given this work's theoretical contention that incivility is multi-faceted in nature, we examined the degree to which the measures correlated with one another when considered as composites. In light of the centrality of these analyses to our ability to address the research question, we supplemented typical hypothesis testing with bootstrapped confidence intervals. These intervals were taken at the 99th percentile, and based on 5,000 bias-corrected (with replacement) re-samples of the data. All statistical analyses were conducted using the *R* statistical computing environment.

Results

Facebook Analytic Sample

As shown in Table 4, the agreement rate between the trace-classified uncivil respondents and self-identified uncivil posters ranged from 72.6% (profane language) to 83.0% (threat). Overall agreement across all item-level indicators was 78.5%. In Table 4, several things stand out. First, across all incivility indicators, there were a small number of instances where a case was classified as uncivil by both the user and the algorithm. Second, the disagreements followed a pattern. For profane language and name-calling, a larger percentage of the disagreements were cases where the user classified themselves as uncivil but the computer classified the user as civil. For the threat and negative stereotypes categories, this pattern was reversed: a greater percentage of the disagreements were in instances where user classified themselves as civil but the algorithm classified the user as uncivil.

[TABLE 4]

Next, examination of the correlations (Table 3 and Figure 3) between the indicator-level variables showed that the self-report measures were all significantly correlated. The self-report indicators had an average inter-item τ value of .39 (r=.54). The trace data incivility indicators

were also all significantly related to each other and had an average inter-item τ value of .36 (r=.54). Looking at the indicator associations across the self-report and trace measures, we found the existence of mostly positive relationships (average inter-item τ = .08 [r=.12]). In contrast to the method-based clusters, not all relationships were statistically significant at p<.05: of the 16 total associations, 5 were significant at p<.05, 4 were marginally significant at p<.10, and 7 were non-significant.

[TABLE 5]
[FIGURE 3]

Examination of the association between the composite-level self-report and trace incivility indicators suggested the existence of a positive and weak relationship, τ =.12, p<.01, $99\%_{CI}$=.03, .21 (r =.19, $99\%_{CI}$=.04, .32).

Twitter Analytic Sample

The agreement rate between self-classified uncivil posters and algorithmically identified uncivil posters (Table 4) ranged from 75.6% (profane language) to 78.2% (threat). The overall agreement rate was 76.9%. Looking at the disagreements, we observed a pattern identical to that found in the Facebook sub-sample: respondents were more likely than the algorithm to classify themselves as uncivil for profane language and name-calling, while the algorithm was more likely than the user to classify the case as uncivil in the instances of threat and use of negative stereotypes.

A τ correlation matrix (Table 6) and accompanying graphical model (Figure 3) indicated that – as was the case for the Facebook analytic sample -- the variables within each method cluster were all significantly related to one another. The average inter-item τ value for the self-report variables was .38 (r=.56) and the average inter-item τ value for the trace variables was .63 (r=.84). Looking across the data collection methods (i.e., examining the relationships between

the self-report and trace indicators), we found that the relationships were all positive in direction. The average inter-item τ value was .09 (r=.14). Not all of these relationships were statistically significant. Specifically, of the 16 associations, 6 were significant at p<.05, 2 were marginally significant at p<.10, and 8 were associated with p-values greater than .10.

[TABLE 6]

Finally, as was the case in the Facebook analytic sample, the association between the indexed measures was positive and significant, τ =.13, p<.01, $99\%_{CI}$=.03, .24 (r =.21, $99\%_{CI}$ =.04,.37).

Discussion

This study set out to assess the degree to that self-reported online political incivility measures corresponded to observed online behavior. The results suggest that a self-reported inventory of online incivility frequency was positively associated with uncivil behavioral instances on both Facebook and Twitter.

Before discussing the implications of these findings, it is important to mention the limitations associated with the current study. Most notably, the observed relationships featured substantial levels of noise. And, as seen in Table 3, the current approach predicted the absence of incivility perhaps better than the presence of incivility. The somewhat weak relationships between the self-report and trace measures of incivility could be due to a number of factors. First, the self-report inventory was not platform-specific. Moreover, the questions were not specific to political communication, while the trace analysis focused specifically on political communication.[8] It seems plausible that a battery of platform and content-specific self-report items may have yielded stronger cross-method correlations. That said, as a proof of concept, it is perhaps in our *favor* that we observed positive and statistically significant relationships between

general measures of self-reported online behavior and more contextually-specific measures of observed online behavior. Second, it's likely the case that both the self-report and trace data measures would benefit from additional, item-level improvement. The self-report measure was designed to somewhat generically capture the most common types of tactical incivility and was not subjected to rigorous empirical refinement. Likewise, the algorithms used to construct the trace measures were selected from an existing battery of annotation tools designed to identify a related, but distinct, social phenomenon (toxicity). Finally, most definitions of incivility involve assessment of communicator intent. For instance, Papacharissi (2004) argued that incivility is a deliberate attempt to use discursive means to undermine democratic functioning. The current study's focus on observed behavior leaves us generally blind to the motivations that animate uncivil behavior on social media.

Despite these limitations, we believe important implications stem from the current data. We found that incivility was relatively rare in occurrence (see Figures 1/2 and Table 3). Both measurement approaches indicated that most people did not habitually engage in uncivil political communication online. When incivility was employed, the results suggested that there exist some congruence between approximated recollections of communication habits and manifest behavior. Localized to the context of online political incivility, our results conform with prior theorizing that suggests incivility is a *conscious act* to disrespect those perceived as oppositional others when conducting political discussion (e.g., Papacharissi, 2004). In other words, the ability of communicators to identify and accurately self-report uncivil communication behaviors provides substantial support for the notion that incivility is not rudeness borne from a spontaneous affective response, but, instead, a deliberate (and perhaps habitual) communication tactic that is centered on the disrespect of others.

Our data also provide some indication that incivility features may cluster together. For example, as seen in Figure 3, the self-reported use of name-calling and the self-reported use negative stereotypes were strongly correlated with each other across both platforms. Likewise, across both platforms, the trace measures of name-calling and invocation of negative stereotypes were robustly associated. Conversely, our data provides some evidence that profane language may not necessarily be deployed in conjunction with other forms of incivility. In other words, within both the self-report and trace-based measures of incivility, profane language -- when compared to the other incivility features -- had comparatively small average inter-item correlation values. This pattern held across both platforms. The idea that users may employ incivility repertoires may help explain some of the low correlations between indicators. These weak relationships may be especially apparent when assessing associations across the self-report and trace measures due to the absence of common method bias. These findings also suggest that incivility might be best understood as a second-order construct that is reflected in multiple first-order variable clusters.

The current data also offer potentially valuable insights as they pertain to human response errors. As shown in Table 3, in the cases of profane language and name-calling, respondents were likely to *over-report* engaging in the behavior (relative to the algorithmic classification). Given that profanity and name-calling are both relatively common social behaviors that are infrequently governed by strong social sanction, it may be the case the discrepancy between reported and observed levels of behavior is related to recall bias, or the failure to accurately recall the degree to which one demonstrates a given behavior. As illustrated by Touangeau (2000), routine behaviors may simply go unnoticed and therefore never become encoded into memory. Alternately, participants were comparatively more likely to *under-report* engaging the

use of threatening language and the use of negative stereotypes. Verbal threats and the use of identity-based hate language are often governed by meaningful social (and sometimes legal) sanction. For this reason, it may be the case that the participant responses were subject to social desirability bias, or the "tendency to present one's self in the best possible light," which can result in the systematic bias of data in the direction of what is "correct or socially acceptable" (Fisher, 1993, p. 303). These findings are important because they suggest that theoretical concepts that are reflected by a variety of behavioral or cognitive features may be internally subject to different sources of error. In this way, computational measures of behavior may be especially valuable because they do not rely on memory-based or socially-influenced reporting process. At the same time, algorithmically-based behavioral measures are subject to the biases present in their human creators and can be also be negatively influenced by sources of systematic internal error (e.g., inability to detect complex forms of social communication). Despite these issues, we believe that studies that draw upon multiple data sources may be key in the development of algorithmic tools that present accurate and comparatively unbiased renderings of human behavior. We should, however, note that the dichotomization process used in Table 3 was rough in nature, and that caution should be taken when interpreting the presented results.

Conclusion

In this study, we identified statistically significant and positive relationships between self-reported and behavioral measures of online political incivility. While these findings are theoretically meaningful, the overarching intent of this work was not necessarily or solely to present a finalized trace measure of online political incivility. Instead, we also sought to use this exploratory investigation as proof of concept that can potentially be used by future researchers to develop valid and reliable measures of online communication behaviors. Based on this

experience, we have five suggestions for future research. Three of these suggestions are specific to the future study of incivility and two of these suggestions relate more generally to work that seeks to combine survey and trace data. As it pertains to the former, our first recommendation is that researchers build upon these findings to continue the work of developing a trace data measure of online political incivility. The results here are a starting, rather than ending, point. Second, we urge researchers to develop a more comprehensive and more rigorously-developed self-report index of online political incivility. It is likely the case that the inventory used here would benefit from additional items and in-depth analysis of the dimensional structure of these items. Third, upon completion of these two tasks, researchers could extend the current methodological approach to better understand the motivations that underlie uncivil political communication. Looking beyond the study of incivility, our first suggestion for future research is to explicitly lay out the costs and benefits of using commercially-developed and publically-accessible computational tools relative to the use of proprietary academic tools developed for specific research problems. On one hand, commercially-developed tools often are the result of resource-intensive developmental efforts. And, their (comparatively) broad accessibility may be attractive to researchers who do not have formal training in the computational social sciences. At the same time, these tools often have to be retrofitted to address specific research questions, and may, therefore, have conceptual limitations. Second, researchers should better define ethical practices for studies that involve the collection of both survey and trace data. As an emergent methodological approach, there do not yet exist best practices guiding critical issues such as consent statement construction and means by which raw replication data can be publically furnished without violating platform terms of service statements.

Biosketches

Toby Hopp is an assistant professor at the University of Colorado Boulder. His research interests are broadly related to the uses and effects of digital and interactive media, the social and motivational factors that underlie uncivil online communication, and organizational transparency.

Chris J. Vargo is an assistant professor specializing in computational social science at the University of Colorado Boulder. He utilizes computer science methods to investigate social media using theories from mass communication and political science disciplines.

Lucas Dixon is Chief Scientist at Google Jigsaw where he works on technologies that help make people safer online. His recent work has been on good conversations online and at scale. This work leverages advances in machine intelligence and explores how it can enable new kinds of user experiences as well as analysis of conversations at scale.

Nithum Thain is a Research Manager at Google Jigsaw where he works on improving online conversation. His focus is on deep learning algorithms for language understanding, interpretable machine learning models, and ML fairness.

Data Availability

Replication data can be obtained from the lead author at tobias.hopp@colorado.edu

References

Brooks, D.J., & Geer, J.G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science, 51,* 1-16. https://doi.org/10.1111/j.1540-5907.2007.00233.x

Chen, G.M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media, 61*, 108–125. https://doi.org/10.1080/08838151.2016.1273922

Coe, K., Kenski, K., & Rains, S.A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64,* 658–679. https://doi.org/10.1111/jcom.12104

Epskamp, S., Borsboom, D., & Fried, E.I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50,* 195-212. https://doi.org/10.3758/s13428-017-0862-1

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*. From http://binf.gmu.edu/mmasso/ROC101.pdf

Fisher, R.J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20,* 303-315. https://doi.org/10.1086/209351

Gervais, B.T. (2015a). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics, 12,* https://doi.org/10.1080/19331681.2014.997416

Gervais, B.T. (2015b). Political incivility online. Retrieved from http://www.ispp.org/jsc/blog/themed-blog-political-incivility-online

Kendall, M. G. (1970). *Rank correlation methods.* London, UK: Charles Griffin

Muddiman, A. (2017). Personal and public levels of incivility. *International Journal of Communication, 11,* 3182–3202.

Mutz, D. (2015). *In-your-face politics: The consequences of uncivil media.* Princeton, NJ: Princeton University Press.

Ng., E.W.J., & Detenber, B.H. (2005). The impact of synchronicity and civility in online political discussions on perceptions and intentions to participate. *Journal of Computer-Mediated Communication, 10*(3). https://doi.org/10.1111/j.1083-6101.2005.tb00252.x

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online discussion groups. *New Media & Society, 6,* 259-283. https://doi.org/10.1177/1461444804041444

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society, 18,* 121–138. https://doi.org/10.1080/1369118X.2014.940365

Santana, A.D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice, 8,* 18–33. https://doi.org/10.1080/17512786.2013.813194

Silva, S., Anunciação, O., & Lotz, M. (2011). A comparison of machine learning methods for the prediction of breast cancer. In C. Pizzuti, M. D. Ritchie, & M. Giacobini (Eds.), *Lecture Notes in Computer Science: Vol. 6623. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 159-170). https://doi.org/10.1007/978-3-642-20389-3_17

Sobieraj, S., & Berry, J.M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication, 28,* 19–41. https://doi.org/10.1080/10584609.2010.542360

Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C.A. Bachrach, J.B. Jobe, H.S. Kurtzman, & V.S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 29-47). Mahwah, NJ: Lawrence Erlbaum Associates.

Vargo, C.J., & Hopp, T. (2017). Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on Twitter: A congressional district-level analysis. *Social Science Computer Review, 35,* 10-32. https://doi.org/10.1177/0894439315602858

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K.L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology, 13,* 1–7. https://doi.org/10.1186/1471-2288-13-61

Notes

1. We instituted an approximate 50/50 gender split because prior experience with the Qualtrics has shown us that samples without gender quotas can result in stark gender imbalances. As it pertains to the active user criteria, we required that all users have a current account with at least 50 pieces of posted content on both platforms. This safeguard was put in place to avoid scenarios where a user may create an account simply to qualify for the study. Of those that engaged with the study materials, 13.5% provided valid data. Participation incentives were provided on the basis of qualifying for and completing the study as a whole (i.e., we did not offer extra incentives for access to social media data).

2. Political talk was coded as being present if a post discussed: legislation/legislative actions; municipal/regional/state political issues; high-profile social issues; elections of voting; the Supreme Court or other high-profile judiciary proceedings with political ramifications.

3. This selected model was an average prediction score from the following models: Random Forests, Gradient Boosted Greedy Trees with early stopping and Kernel SVM classifiers. Ensemble models can deliver superior classification due to their ability to leverage multiple machine learning models at once.

4. For API documentation visit https://github.com/conversationai/perspectiveapi

5. Given the low prevalence of uncivil posts, we report Gwet's AC1 rather than more commonly reported reliability measures (e.g., Cohen's Kappa). When feature prevalence is very low, Cohen's Kappa (and similar indexes) have been shown to result in downwardly-biased reliability estimates (Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013).

6. To further delineate the need for weighting, consider Respondent A who posts a single, very uncivil political comment (probability score =1.00). Contrast this with Respondent B, who posts 20 political comments containing high levels of incivility (average probability score =.95). If we simply employed user-level averages, Respondent A would be assessed as more uncivil than Respondent B, despite the fact that Respondent B likely exerts a more toxic influence on her/his discussion environment(s). In an even more extreme case, compare Respondent A to Respondent C, who posts 100 political comments all of which are assigned an incivility score of 1. The simple application of user-level averages would have resulted in the classification of Respondents A and B as equally uncivil.

7. A total of 204 participants created at least one political post on both platforms.

8. Participants were not required to actively engage in political discussion to be included in the sample. Thus, to preserve face validity of the questions, we oriented them towards online communication in general.

Table 1

Description of the Perspective API Components Used to Construct the Trace Data Measure of Incivility

Incivility Dimension	Perspective API Component	Description
Profane Language	Obscene Language	Swear words or other vulgar, explicit or offensive language
Name-Calling	Insulting Language	Derogatory name calling or putting others down
Threat	Threatening Language	Verbal intention to inflict pain, injury, or violence against an individual or group
Invocation of Negative Stereotypes	Hateful Language	Anger, disgust, hatred, other negative emotions against a person or group based on identity attributes

Table 2

Example Showing Application of Weight Value in the Computation of Final Trace Indicator Scores

	Low Profane Language (0.004)	**Moderate Profane Language (0.19)**	**High Profane Language (0.99)**
Infrequent Online Political Communication (*w* =0.08)	0.0003	0.02	0.08
Moderate Online Political Communication (*w* =0.74)	0.003	0.14	0.73
Frequent Online Political Communication (*w* =5.42)	0.02	1.03	5.37

Note. For the purposes of illustration, this table uses the raw indicator score for the Facebook-based measure of profane language. Categories represent minimum, mean, and max observed values for the *w* and profane language variables. As shown, the *w* variable functions in such a way that the final computed score is highest for frequent posters who are also high in profane language, while lower scores are assigned to those who avoid the use of profane language or post political content infrequently

Table 3

Descriptive Statistics for Full Dataset and Analytic Samples

	Full Dataset	Facebook Analytic Sample	Twitter Analytic Sample
N	783	442	270
$M_{\text{Incivility Self-Report Composite}}$ (SD)	2.04(1.17)	2.09(1.10)	2.13(1.15)
$\alpha_{\text{Incivility Self-Report Composite}}$	.80	.77	.77
$M_{\text{Profane Language Self-Report}}$ (SD)	2.65(1.84)	2.67(1.82)	2.67(1.80)
$M_{\text{Name-Calling Self-Report}}$ (SD)	2.37(1.67)	2.50(1.63)	2.58(1.73)
$M_{\text{Threat Self-Report}}$ (SD)	1.36(1.08)	1.30(0.97)	1.38(1.12)
$M_{\text{Invocation of Negative Stereotypes Self-Report}}$ (SD)	1.79(1.29)	1.88(1.28)	1.90(1.29)
$M_{\text{Incivility Trace Composite [Facebook]}}$ (SD)		0.15(0.15)	
$\alpha_{\text{Incivility Trace Composite [Facebook]}}$		.69	
$M_{\text{Profane Language Trace [Facebook]}}$ (SD)		0.19(0.20)	
$M_{\text{Name-Calling Trace [Facebook]}}$ (SD)		0.09(0.13)	
$M_{\text{Threat Trace [Facebook]}}$ (SD)		0.07(0.11)	
$M_{\text{Invocation of Negative Stereotypes Trace [Facebook]}}$ (SD)		0.25(0.35)	
$M_{\text{Incivility Trace [Twitter]}}$ (SD)			0.15(0.17)
$\alpha_{\text{Incivility Trace [Twitter]}}$			.86
$M_{\text{Profane Language Trace [Twitter]}}$ (SD)			0.11(0.13)
$M_{\text{Name-Calling Trace [Twitter]}}$ (SD)			0.17(0.23)
$M_{\text{Threat Trace [Twitter]}}$ (SD)			0.13(0.16)
$M_{\text{Invocation of Negative Stereotypes Trace [Twitter]}}$ (SD)			0.19(0.24)
% Male	41.7%	46.0%	53.5%
M_{age} (SD)	39.32(12.90)	41.22(13.10)	41.26(14.16)
$M_{\text{Conservatism}}$ (SD) [1 = very liberal, 7 = very conservative]	3.72(1.85)	3.59(1.92)	3.43(2.00)
% Democrat	39.9%	40.8%	45.4%
% Republican	23.7%	22.0%	21.9%
% Independent	32.6%	32.2%	29.4%
% Other	3.8%	5.0%	3.4%
% Vote 2016	83.5%	86.4%	88.9%
$M_{\text{Facebook Intensity}}$ (SD) [1 = use infrequently, 7 = use frequently]	6.46(1.09)	6.54(0.99)	6.38(1.18)
$M_{\text{Twitter Intensity}}$ (SD) [1 = use infrequently, 7 = use frequently]	5.22(1.74)	5.19(1.75)	5.61(1.56)
$M_{\text{Number of Facebook Political Posts}}$ (SD)		22.56(91.83)	
$M_{\text{Number of Twitter Political Posts}}$ (SD)			21.89(96.81)
$M_{\text{Facebook Activity Years}}$ (SD)		7.52(2.30)	7.35(2.49)
$M_{\text{Twitter Activity Years}}$ (SD)		4.11(2.59)	4.01(2.61)

Table 4

Agreement Between Self-reported and Behaviorally Identified Uncivil Respondents

Facebook Analytic Sample (*n* = 442)

	Profane Language		Name-Calling		Threat		Invocation of Negative Stereotypes	
	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)
Civil (Self-Report)	313	44	339	42	363	46	341	49
Uncivil (Self-Report)	77	8	53	8	29	4	41	11
% Agreement	**72.6%**		**73.8%**		**83.0%**		**79.6%**	

Twitter Analytic Sample (*n* = 270)

	Profane Language		Name-Calling		Threat		Invocation of Negative Stereotypes	
	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)	Civil (Trace)	Uncivil (Trace)
Civil (Self-Report)	194	25	198	30	207	39	202	34
Uncivil (Self-Report)	41	10	34	10	20	4	29	5
% Agreement	**75.6%**		**77.0%**		**78.2%**		**76.7%**	

Note. Cases were coded as uncivil if scores were more than 1 standard deviation above the indicator's mean value. % agreement was calculated as the number of times the self-report and trace measures corresponded taken over the subsample *n*. Overall % agreement for Facebook = 78.5% (1387/1768). Overall % agreement for Twitter = 76.9% (830/1080).

Table 5

Correlations Between Self-report and Trace Data Measures of Incivility (Facebook Analytic Subsample)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Profane Language (Self-Report)**[1]**	--	**.46**	**.34**	**.31**	**.11**	-.03	.09†	.04
Name-Calling (Self-Report)**[2]**	**.66**	--	**.36**	**.48**	.07	.06	**.11**	**.11**
Threat (Self-Report)**[3]**	**.51**	**.54**	--	**.37**	.00	-.03	.08†	.03
Invocation of Negative Stereotypes (Self-Report)**[4]**	**.47**	**.68**	**.55**	--	.08†	.09†	**.21**	**.21**
Profane Language (Trace)**[5]**	**.18**	.11	.01	.13	--	**.20**	**.32**	**.25**
Name-Calling (Trace)**[6]**	-.05	.10	-.04	.14	**.30**	--	**.33**	**.43**
Threat (Trace)**[7]**	.14	**.18**	.13	**.33**	**.48**	**.49**	--	**.64**
Invocation of Negative Stereotypes (Trace)**[8]**	**.06**	**.17**	.05	**.32**	**.39**	**.62**	**.84**	--

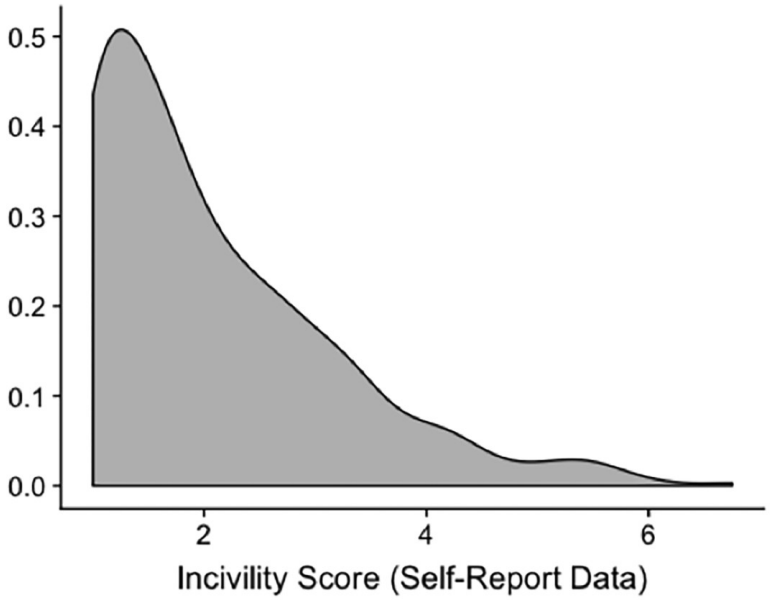
Note. Kendall's τ above the diagonal, Pearson's *r* equivalencies below the diagonal; Bolded coefficients significant at $p<.05$; Coefficients marked with a † significant at $p<.10$

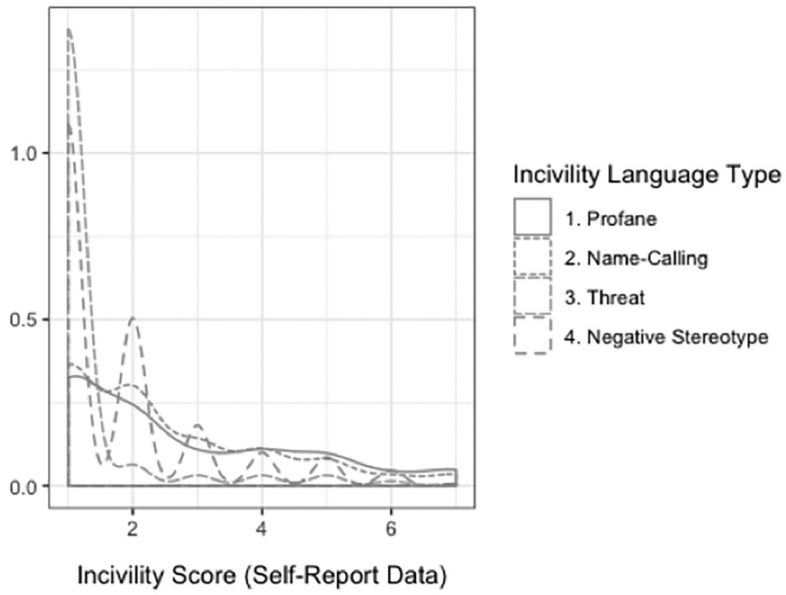
Table 6

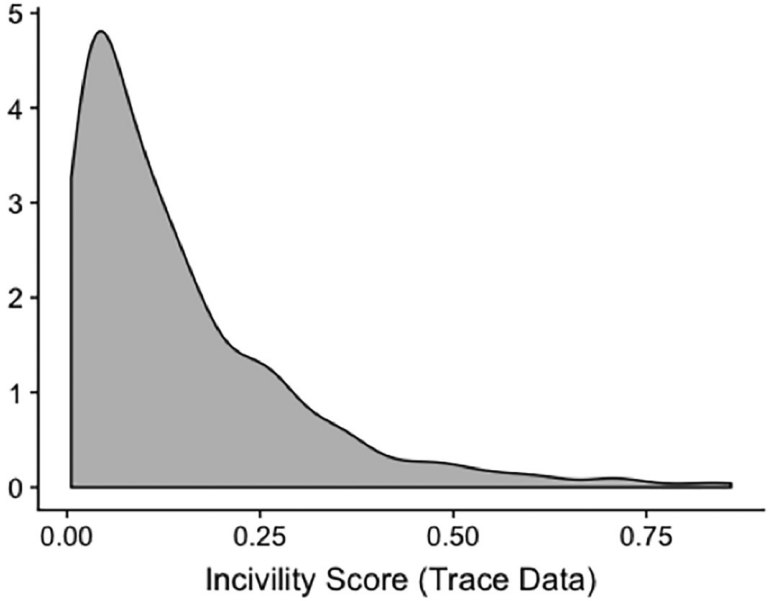
Correlations Between Self-report and Trace Data Measures of Incivility (Twitter Analytic Subsample)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Profane Language (Self-Report)**[1]**	--	**.40**	**.36**	**.24**	**.12**	.07	.08	.07
Name-Calling (Self-Report)**[2]**	**.59**	--	**.39**	**.48**	.12†	**.15**	**.13**	**.14**
Threat (Self-Report)**[3]**	**.53**	**.58**	--	**.40**	.04	.02	.02	.02
Invocation of Negative Stereotypes (Self-Report)**[4]**	**.37**	**.69**	**.59**	--	.08	**.13**	**.13**	.12†
Profane Language (Trace)**[5]**	**.19**	.18	.06	.13	--	**.47**	**.41**	**.44**
Name-Calling (Trace)**[6]**	.12	**.24**	.04	**.21**	**.68**	--	**.77**	**.90**
Threat (Trace)**[7]**	.13	**.21**	.04	**.20**	**.61**	**.94**	--	**.80**
Invocation of Negative Stereotypes (Trace)**[8]**	.11	**.21**	.03	.18	**.63**	**.99**	**.95**	--

Note. Kendall's τ above the diagonal, Pearson's *r* equivalencies below the diagonal; Bolded coefficients significant at $p<.05$; Coefficients marked with a † significant at $p<.10$


0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
Incivility Score (Self-Report Data)


1.0
0.5
0.0
2
4
6
Incivility Score (Self-Report Data)
Incivility Language Type
1. Profane
2. Name-Calling
3. Threat
4. Negative Stereotype


5
4
3
2
1
0
0.00
0.25
0.50
0.75
Incivility Score (Trace Data)

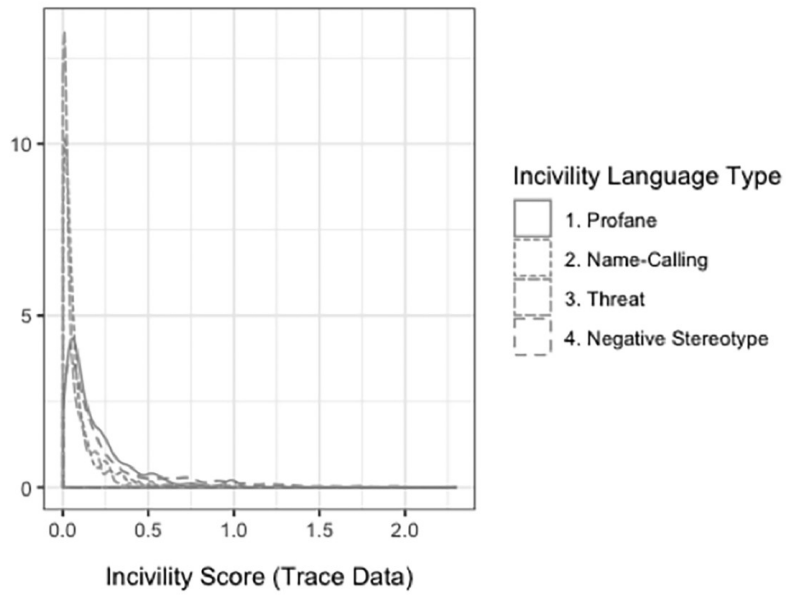
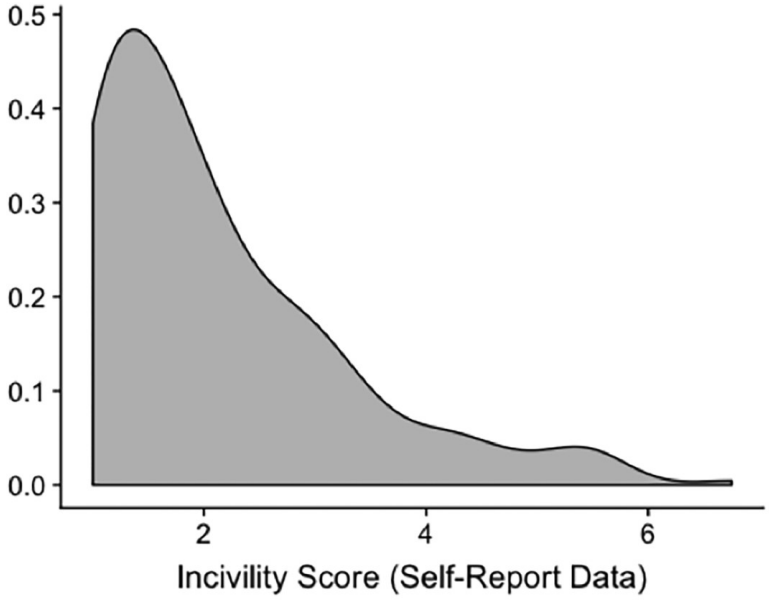
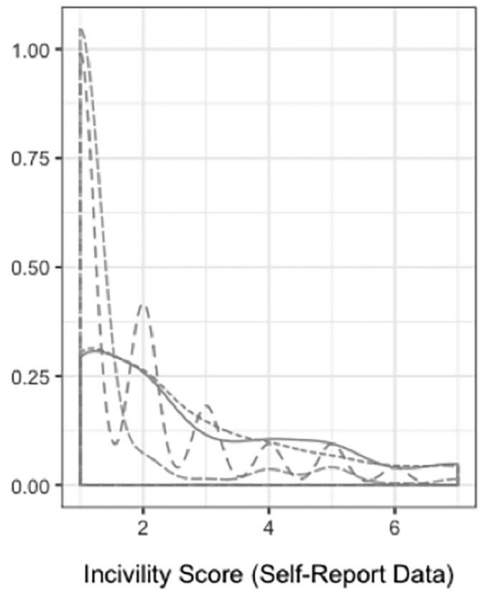
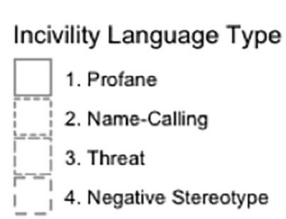
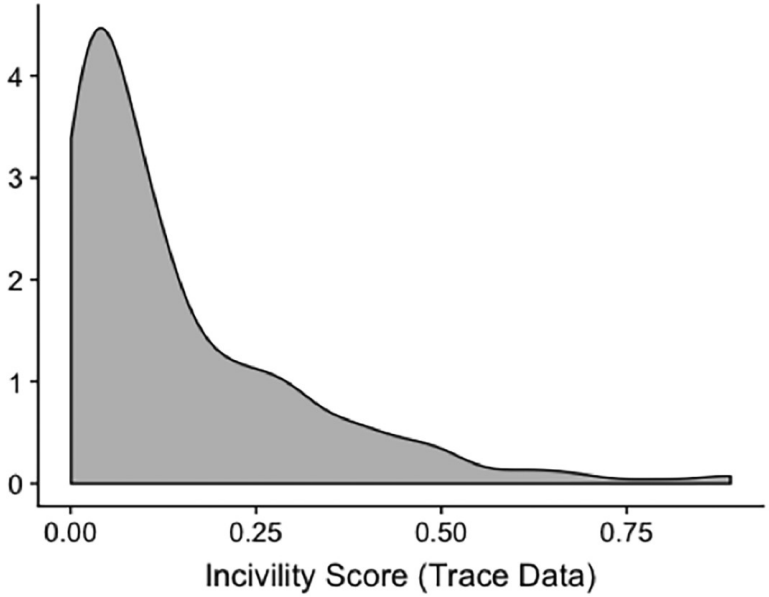

10
5
0
0.0
0.5
1.0
1.5
2.0
Incivility Score (Trace Data)
Incivility Language Type
1. Profane
2. Name-Calling
3. Threat
4. Negative Stereotype

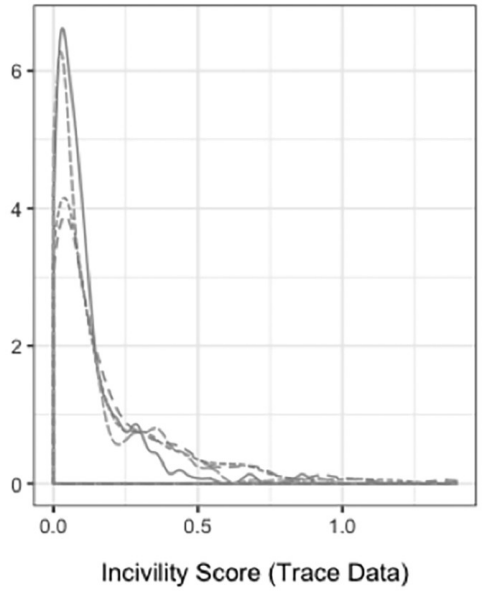
Figure 1. Density plots for incivility indicator variables (Facebook analytic sample)
Note: The upper left plot shows the density for the composite self-report incivility variable. The upper right plot shows the density for the individual indicators comprising the composite variable. This pattern is repeated in the lower left and right panels for the trace incivility variables.


0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
Incivility Score (Self-Report Data)


1.00
0.75
0.50
0.25
0.00
2
4
6
Incivility Score (Self-Report Data)


Incivility Language Type
1. Profane
2. Name-Calling
3. Threat
4. Negative Stereotype


4
3
2
1
0
0.00
0.25
0.50
0.75
Incivility Score (Trace Data)


6
4
2
0
0.0
0.5
1.0
Incivility Score (Trace Data)

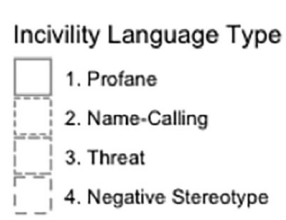
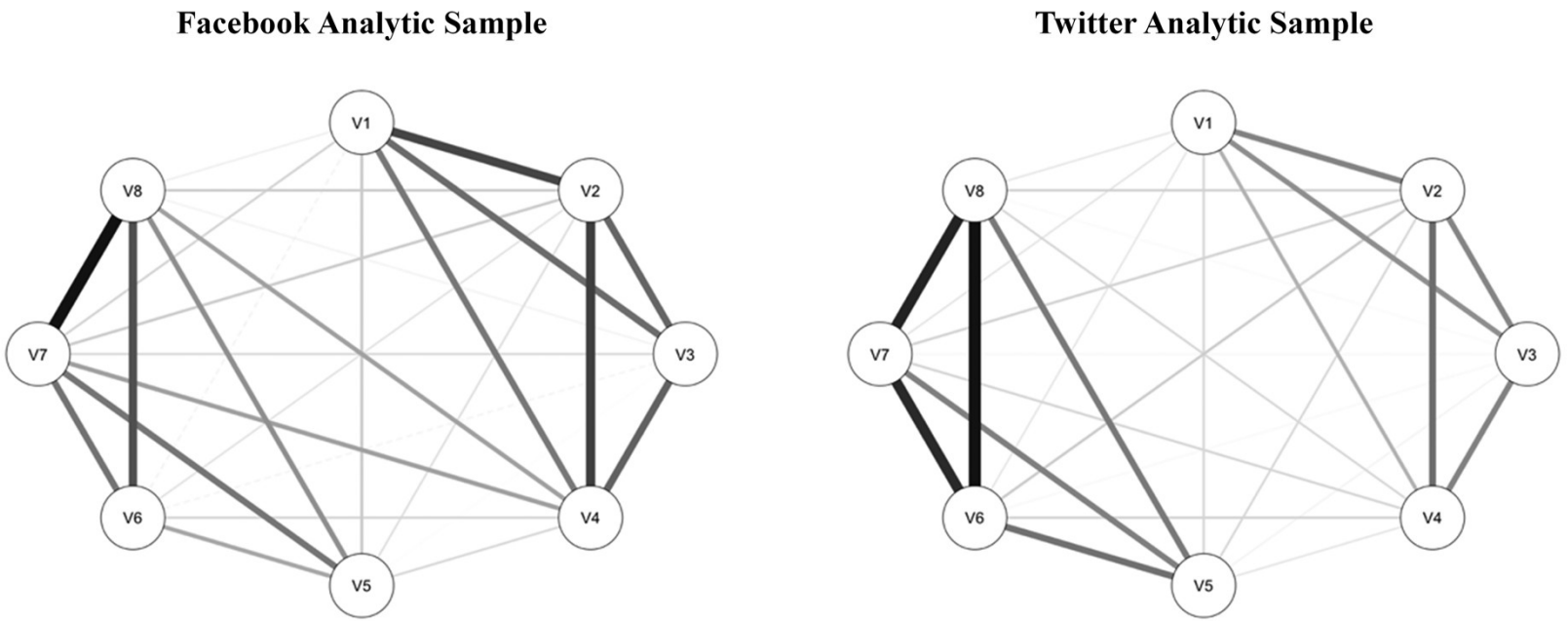

Incivility Language Type
1. Profane
2. Name-Calling
3. Threat
4. Negative Stereotype

Figure 2. Density plots for incivility indicator variables (Twitter analytic sample)
Note: The upper left plot shows the density for the composite self-report incivility variable. The upper right plot shows the density for the individual indicators comprising the composite variable. This pattern is repeated in the lower left and right panels for the trace incivility variables.



Key
V1 Profane Language (Self-Report)
V2Name-Calling (Self-Report)
V3 Threat (Self-Report)
V4 Invocation of Negative Stereotypes (Self-Report)
V5 Profane Language (Trace)
V6 Name-Calling (Trace)
V7 Threat (Trace)
V8 Invocation of Negative Stereotypes (Trace)

Note: Thicker lines indicate stronger relationships. Solid/dotted lines indicate positive/negative relationships.