Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter

Kevin Munger*

May 20, 2020

Abstract

I conduct an experiment which examines the impact of moral suasion on partisans engaged in uncivil arguments. Partisans often respond in vitriolic ways to politicians they disagree with, and this can engender hateful responses from partisans from the other side. This phenomenon was especially common during the contentious 2016 US Presidential Election. Using Twitter accounts that I controlled, I sanctioned people engaged partisan incivility in October 2016. I found that messages containing moral suasion were more effective at reducing incivility than were messages with no moral content in the first week post-treatment. There were no significant treatment effects in the first day post-treatment, emphasizing the need for research designs that measure effect duration. The type of moral suasion employed, however, did not have the expected differential effect

*I would like to thank many people for their feedback on this paper, especially the members of the NYU SMaPP lab, Livio D. Lonardo, J. Hodgdon Bisbee, Drew Dimmery, Neal Beck, Josh Tucker, Jonathan Nagler, Patrick Egan, Chris Dawes, Andy Guess, Alex Siegel, Joanna Sterling, John Jost, and participants at the several conferences I presented this work in during 2017. I declare no conflicts of interest. The data, code, and any additional materials required to replicate all analyses in this article are available at the Journal of Experimental Political Science Dataverse within the Harvard Dataverse Network, at: doi:10.7910/DVN/OUYTUP.

on either Republicans or Democrats. These effects were significantly moderated by the anonymity of the subjects.

In October of 2016, President Obama claimed (and Democratic presidential nominee Hillary Clinton tweeted) that "civility is on the ballot." Concern over political civility was widespread during the 2016 US presidential election, and many felt that the internet and social media (which Republican presidential nominee Donald Trump employed enthusiastically) were to blame.

Concern over incivility in contemporary political discourse can be traced back at least to the rise of cable news and its personalistic, outraged style (Mutz, 2015; Berry and Sobieraj, 2013). Indeed, concern about civil discourse accompanies any technological advance that lowers the cost of information production and distribution. The invention of the printing press led to elite concern about civil discourse during the time of the Reformation (Bejan, 2017).

Modern technological changes are taking place in the context of increased partisan animosity. Often called "affective polarization," this animosity reflects a growing distrust and lack of respect among Democrats and Republicans (Iyengar, Sood, and Lelkes, 2012). This phenomenon is directly related to civility, which Mutz (2015) says is "a means of demonstrating mutual respect" (2015, p7). Incivility is more than impoliteness: it is indicative of a disregard for the act of deliberation. Internet technologies may or may not be driving affective polarization, but they do at a minimum allow for the lack of mutual respect to manifest itself in uncivil online discourse.

Online communication lacks the evolved social and emotional feedback mechanisms that make it difficult to be uncivil in a real-world setting, and it affords physical distance and (sometimes) anonymity, decreasing the effectiveness of social sanctioning (Frijda, 1988). These technological affordances, in a context replete with bad actors intent on sowing discord for fun (Phillips, 2015) or geopolitical advantage (Chen, 2015), have degraded norms of civil discourse online.

The implications of these changing norms are serious. Early enthusiasm about

the capacity of the internet to democratize political discourse may have been premature (Hindman, 2008), but the affordances of today's ubiquitous, easy-to-use and social internet have caught up with the hype: in 2008, only 25% of US adults were on a social network, but during the 2016 US presidential campaign, that number was 68% (Greenwood, Perrin, and Duggan, 2016).

However, survey evidence suggests that internet users do not feel that their interactions are civil or productive—"64% say their online encounters with people on the opposite side of the political spectrum leave them feeling as if they have even less in common than they thought" (Duggan and Smith, 2016).

I conducted an experiment that evaluates different strategies for promoting civil political discourse during the 2016 US presidential election. Using the method developed by Munger (2017), I used Twitter accounts that I controlled to sanction users engaged in uncivil discussions. In contrast to lab experiments conducted on a convenience sample in a short time frame, this approach allowed me to measure the effectiveness of sanctioning on a sample of frequently uncivil partisans in a realistic setting and in a continuous and unbounded time frame.[1]

By manipulating the partisan identity of my "bots,"[2] I test the differential effects of sanctioning on Republicans and Democrats. By varying the language I tweeted at subjects, I test hypotheses about the relative effectiveness of two kinds of moral suasion and include a non-moral message that simply reminds subjects that what they are tweeting is public.[3]

I found evidence of significant changes in subjects' behavior, but the treatment effects were not what I theorized. Although one moral message caused a significant

[1]All research activities received prior approval by [Home Institution's] Internal Review Board.

[2]These are not "bots" in the sense that they behave autonomously. I did all of the tweeting manually. I refer to them as bots throughout the paper for lack of a better term.

[3]The research design, dependent variable measurement, and main hypothesis were pre-registered at EGAP.org prior to any research activities.

reduction in subject incivility in the first week after treatment and the other did not, these two treatment effects were not significantly different from each other. Furthermore, the theorized differential treatment effect on Democrat and Republican subjects were not observed, although the limited sample size means that this is evidence only against very large treatment effect heterogeneities.

Subject anonymity significantly moderated treatment effects, in the expected direction: more anonymous subjects (defined as the amount of identifying information provided on their Twitter accounts) were less likely to respond to the treatment. This trend was only significant in the one-day time frame.

I also theorized that both moral treatments would be more effective at reducing incivility than a non-moral message that reminded users that what they were saying was public. I found that the moral messages caused a significant reduction, and that this effect was significantly larger than the effect of the non-moral message.

These findings demonstrate that various different forms of moral suasion can be effective in promoting a more civil political discourse on Twitter, above and beyond the effect of merely calling attention to the subjects' behavior. This moral suasion may only be effective on a subset of users; anonymous users (those more likely to be trolls) were unresponsive to moral suasion, and may even have been encouraged by being told that they were violating norms of political civility. Efforts to promote online civility should be sure to target the right people and use the most appropriate rhetorical strategy to maximize their efficacy.

1 Experimentally Reducing Political Incivility

The first step in performing this experiment was finding conversations that were uncivil, between out-partisans, *and* about politics. I thus used streamR, an R package developed

to scrape the streaming Twitter API (Barberá, 2014), to find tweets mentioning either "@realDonaldTrump" or "@HillaryClinton"—the Twitter accounts of the two major party candidates in the 2016 US presidential election. I then dropped any tweets that were not directed at another user who was *not* either Trump or Clinton.

In this way, I found a sample of tweets from non-elites that were concerned with the "issues" most likely to inspire political incivility in October 2016: Trump and Clinton. In order to filter through the hundreds of thousands of tweets every hour that fit these criteria, I used a machine learning classifier designed to detect aggression, assigning an "aggression score" to each tweet I had scraped, then manually evaluated the top 10% most aggressive tweets per batch. Details of this measure are in the following section.

The first step was to check whether the uncivil language was directed at an account that appeared to be a member of the opposite political persuasion. Many of the potential subjects I found this way were tweeting at elites–either people verified on Twitter, journalists or campaign operatives–and I excluded them.

When performing the manual inspection of the potential subject's profile, I excluded users who appeared to be minors or who were not tweeting in English. I also checked to ensure that the subject's profile was at least two months old; Twitter does ban some user accounts for harassment or other violations of their Terms of Service, so a very new account is likely to have been started by someone who had previously been banned.

For a visual overview of this selection process, see Figure 2. In this way, I found uncivil tweets from a non-elite to another non-elite with whom they disagreed politically. For an example, see Figure 1. @realDonaldTrump tweeted something, then Parker tweeted "you already lost" at Trump. Ty then responded to Parker (but because of how Twitter works, Ty's tweet also "mentions" @realDonaldTrump) with an uncivil comment. Ty is the subject I included in the experiment, and because he was being uncivil to someone criticizing Trump, I coded Ty as a Trump supporter.

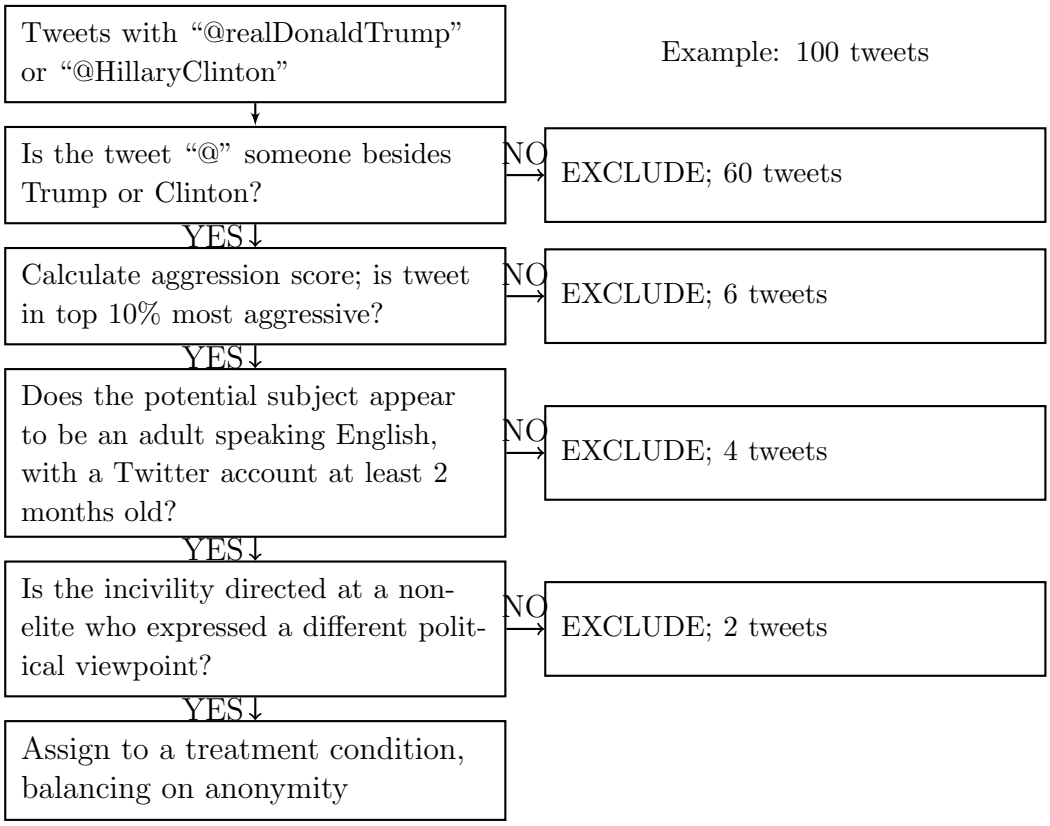
Figure 1: Finding Non-Elite Incivility

The sample in this experiment is thus not a "representative sample" of the general population, or even of Twitter users. The group of people encountered by following the steps listed in Figure 2 are, however, precisely the type of Twitter users who might be able to deliberate with and learn from their political opponents, if they were to do so in a civil fashion.

In the time since I performed this experiment, the prevalence of bots and Russian-backed trolls on Twitter has been revealed. Increasingly sophisticated "social spam bots" have been developed that are difficult for human coders to detect (Cresci et al., 2017). It is eminently possible that some of the subjects in my sample were undetected spam bots or Russian-backed trolls. Furthermore, based on the evidence showing that the majority of Russian bots were posing as Trump supporters, there may have been more of them in the Republican sample of my experiment. However, because my treatments were randomly assigned, there is no reason to think that the prevalence of illegitimate subjects varied across treatment conditions, so they should not pose a problem for inference.

I also recorded each subject's Anonymity Score during the subject discovery process. The Anonymity Score ranged from 0 (least anonymous, full name and picture) to 2

Figure 2: Sample Selection Process



The decision process by which potential subjects were discovered, vetted and ultimately included or excluded. The right-hand column gives an example of the number of potential subjects excluded at each step in the process.

(most anonymous, no identifying information). Ty, from Figure 1, was coded as a 1—he chose to display what could plausibly be his full name.

My aim was to convince subjects that they were being sanctioned by a real person, so I made my bots look as real as possible. After I tweeted at a subject, they were likely to have received a "notification" from Twitter.[4] It is uncommon to be tweeted at by a stranger, but not extremely so, and especially not among a subject pool who are tweeting uncivil things at out-partisans. As a result, they were likely to click on my bots' profile; if they did, they would see something very like Figure 3.

Neil, in panel (a), was a bot who appeared to be pro-Clinton. I created four bots; the other three were pro-Democrats, pro-Trump, and pro-Republicans (see Todd, in panel (b)). To manipulate these identities, I changed the large banner in the middle of the profile, the small logo in the bottom right of the bots' profile pictures, and the "bio" field below their username (eg "Hillary 2016!"; "Republicans 2016!"). The four bots were otherwise identical. All of the bots appeared to be white men, keeping the race/gender aspect of the treatment constant.

I took other steps in order to maximize verisimilitude. Most importantly, I ensured that all of the bots had a reasonably high number of followers. Munger (2017) varied the number of followers that sanctioning bots had, and found that bots with few followers had very little effect. Based on this finding, I purchased 900 followers for each.

I created each bot in January 2015, giving the impression that they were long-time users.To further increase the perception that the bot was a real person, I tweeted dozens of innocuous observations and retweeted random (non-political) stories from the accounts the bots followed.

There were two subject pools: people who were uncivil to people critical of Trump ("Republicans") and people who were uncivil to people critical of Clinton ("Democrats").

[4]If subjects turned off notifications, some of them might not have received the treatment, which should bias my treatment effects towards zero.

Figure 3: (a) Example Bot–Clinton Condition

(b) Example Bot–Republican Condition

Within each of these pools, each subject was randomly assigned one of three messages ("Feelings", "Rules", or "Public") sent by one of two bots (pro-candidate or pro-party). There were initially 118 subjects in the "Republicans" pool, 104 subjects in the "Democrats" pool, and another 108 in the control group, to whom I sent no tweets.[5]

The primary variation in the treatments was in the language of the message sent to the subjects. The aim was to convince subjects to change their behavior. One approach, employed in Munger (2017), is *in-group social norm promotion*: to cause subjects to update their beliefs about normative behavior for someone sharing their social identity. Munger found that sanctioning from bots that shared a social identity with the subject was more effective in changing their behavior than bots with a different social identity. To build on this finding, I held in-group social identity (in this case, partisanship) constant in the current study.

By varying the language of the in-group sanctioning, I tested the possibility of moral suasion. I based my approach on the moral intuitionist model proposed by Haidt (2001), which argues that moral emotion is antecedent to moral reasoning. Extending the theory, (Haidt, 2012) argues that a necessary component for moral suasion is convincing your interlocutor that you are sympathetic and understanding. If the two of you share the same fundamental moral intuitions, you can reasonably discuss specific implications of those foundations, but if not, attempts to change their mind are likely to be interpreted as attacks on their worldview and to be met with resistance.

I designed two different treatments. The first was designed to appeal to the Care foundation, and thus to have some effect on Republicans but a much larger effect on Democrats[6]:

[5]In the analysis below, I include 310 subjects out of this original pool of 330. I discuss the attrition process in Appendix A.

[6]In *The Righteous Mind*, Haidt argues that Democrats' morality is built on Care, but specifically on care for certain victim groups who have traditionally been marginalized in US society. This treatment

```
@[subject] You shouldn't use language like that. [Republicans/Democrats]
 need to remember that our opponents are real people, with real feelings.
```

The other treatment appealed to the Authority foundation. My expectation was that it should have an effect on Republicans but not on Democrats:

```
@[subject] You shouldn't use language like that. [Republicans/Democrats]
need to behave according to the proper rules of political civility.
```

In addition to these moral foundations treatments, I included a non-moral "public" treatment. My intention was to use a message that would serve to remind subjects that their uncivil tweets were public, and my hypothesis was that this treatment would decrease the subjects' use of incivility, but that the effect would be smaller than the moral treatments'. To that end, I designed a message that emphasized the subject's visibility:

```
@[subject] Remember that everything you post here is public. Everyone
 can see that you tweeted this.
```

Hypothesis 1 *The reduction in incivility caused by the Care condition will be larger for Democrats than for Republicans. There should be a reduction in incivility caused by the Authority condition for Republicans, but not for Democrats. There should be a reduction in incivility cased by the Public condition, but it should be smaller than the other effects.*

Some subjects are more heavily invested in their online identities than are others. Twitter allows individuals to decide how much personal information to divulge, so while some users are completely anonymous, others include their full name, picture,

would thus be less effective if Democrat subjects perceive their Republican interlocutors to not be deserving of care.

and biography. Users who are more invested in their online identities are more likely to change their behavior in response to sanctioning, while anonymous users are unlikely to do so.[7]

Hypothesis 2 *The reduction in incivility caused by the treatments will positively covary with the subject's Anonymity Score.*

2 Results

The behavior targeted in this experiment is partisan incivility targeted at other Twitter users. To capture this behavior, I scraped each subject's Twitter history before and after the treatment and restricted the sample to the tweets that were "@-replies": tweets directed at another user. After removing the 18 users for whom I could not collect enough pre- or post-treatment tweets (see Appendix A for a full discussion), I used the model trained by Wulczyn, Thain, and Dixon (2017) to assign an "aggression score" (between 0 and 1) to each of these 367 thousand tweets. This measure was skewed toward the lower end of the distribution, so I selected all tweets above the 70th percentile aggression score and coded them as incivil.[8]

I selected the 70th percentile based on the empirical distribution of aggression scores and the concordance with the validation data discussed below. See Appendix B for the empirical distribution. This exact cutoff was not specified in my pre-analysis plan, but the fact that I would be using this model was pre-registered.

Aggression is far from the only measure of incivility that scholars have proposed. Note, however, that the current model measures aggression in a set of subjects who have been aggressive to non-elite members of the out-party, making it similar to the

[7]Note that this hypothesis was not recorded in the Pre-Analysis Plan, but follows directly from the theory in Munger (2017).

[8]Appendix D presents a robustness check of this threshold. For the full data and replication analysis, see Munger (2020)

definition of incivility used by Theocharis et al. (2015). Perhaps the most similar operationalization is the "personal-level incivility" defined by Muddiman (2017), who find that this form of incivility is evaluated by subjects on Mechanical Turk to be more uncivil than other conceptions of incivility.

This result mirrors my own validation of the Wikipedia model. I had a random sample of 1,000 subject tweets labeled as either "civil" or "uncivil" by crowdworkers on Mechanical Turk. The model correctly predicted the human-generated labels 81.2% of the time while ensuring a balance between type 1 and type 2 classification error (for a more extensive discussion, see Appendix C).

The aggression model is theoretically consonant with others in the literature, but equally importantly (in my view), it was specified in my pre-registration of this research. This pre-registration is particularly important when using measures derived from text. Text data is extremely high dimensional, so the development of measures *ex post* allows researchers (often unknowingly) to select the measure out of millions of potential measures that best supports their hypothesis.

To control for each subject's pre-treatment behavior, I calculated their rate of uncivil tweeting in the three months before the experiment. This measure was included as a covariate in all of the following analysis.

The data take the form of overdispersed count data: the variables that record the number of uncivil tweets sent by each user are bounded by zero and vary widely between highly active and normal users. Table 1 reports these distributions. To account for this high variance, I take the log of the uncivil tweet count variables in the following results.[9]

The experimental results on the full sample with all treatments pooled are displayed in Figure 4. In all of the analysis that follows, the dependent variable is the (log of the)

[9]Another approach to handling overdispersed count data is to fit a negative binomial model. The results of this model can be found in Appendix E. Another approach would be to divide subjects by how often they tweeted, to see if treatment effects are constant across subject loquacity. Appendix F shows that this is not the case

Table 1: Distribution of Incivil Subject Tweets, Pre- and Post-Treatment

	1st Quartile	Median	3rd Quartile	Mean
Pre-Treatment (90 days)	124	349	790	570
Pre-Treatment uncivil	36	105	257	171
Post-Treatment (43 days)	110	322	786	609
Post-Treatment uncivil	32	97	235	170
Republicans: pre-Treatment uncivil	36	97	231	151
Democrats: pre-Treatment uncivil	36	125	292	193

number of uncivil tweets the subject sent in the specified time period. Each of the four results are for a given non-overlapping time period: the first result is for the first day after treatment, the second result for days 2-7, the third for days 8-14, and the fourth for days 15-28.

In none of the these four time periods is there a statistically significant treatment effect with all three treatments pooled. To test the second part of Hypothesis 1 (that the effect of the two moral treatments would be larger than the Public treatment), Figure 5 pools the three treatments into these two categories. In Week 1, there is a statistically significant effect of the two moral treatments, while the effect of the Public treatment is actually slightly positive. These treatment effects are statistically significantly different from each other ($p < .02$).

Although not an explicit hypothesis, it is plausible that treatment effects should decay over time. This does not seem to be the case in Figure 5: there are no significant treatment effects in the first day after treatment, but there are in the first week after treatment. This is to some extent an issue of data size: the first twenty-four hour time period is simply too noisy. The results in Appendix D support this conclusion: if the analysis is run with the higher threshold for classifying tweets as uncivil (thus decreasing the number of tweets in the analysis), the point estimate of the effect in Day 1 becomes significant for the moral treatments.

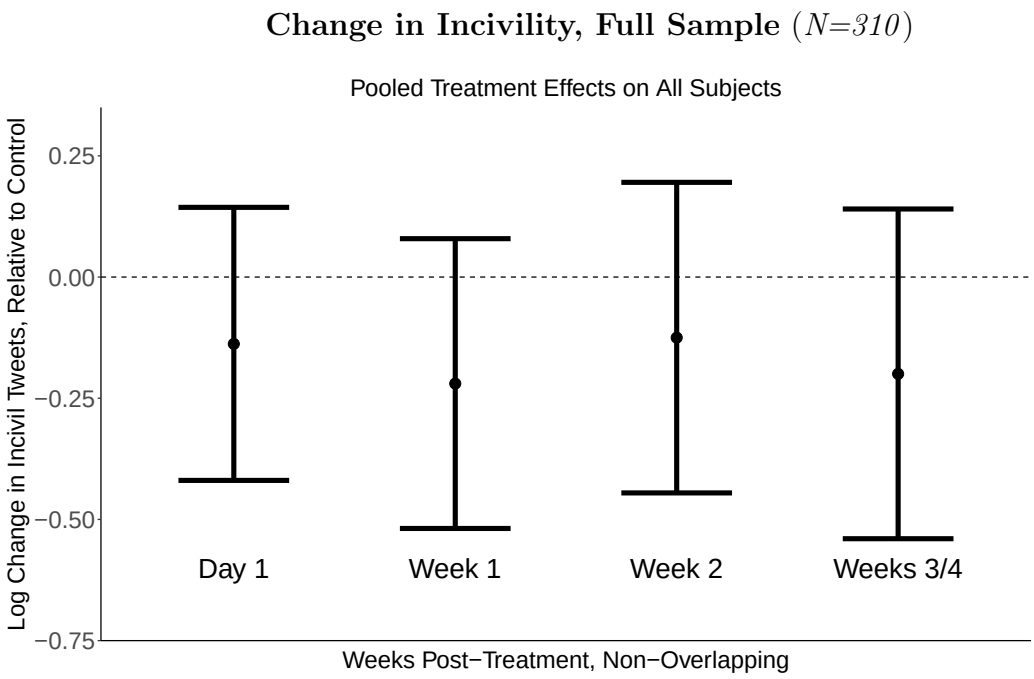


Figure 4: Pooled treatment effects on the entire sample, controlling for the log of the number of pre-treatment uncivil tweets sent by each subject. Lines represent 95% confidence intervals.

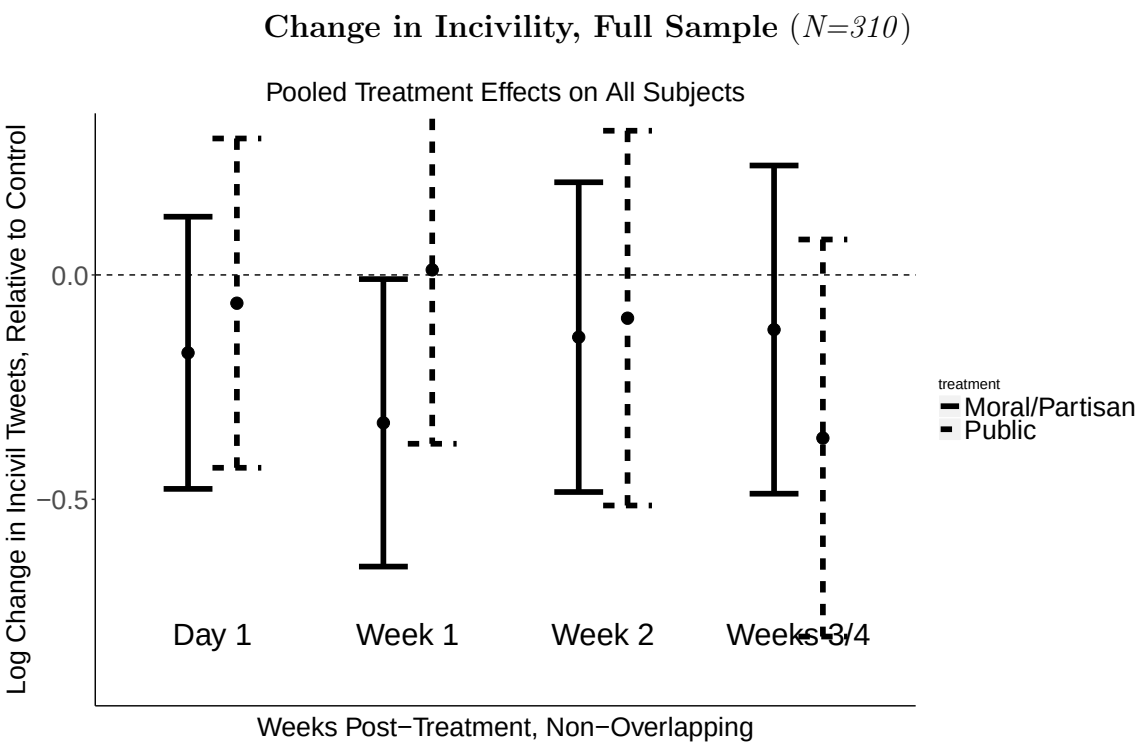


Figure 5: Pooled treatment effects on the entire sample.

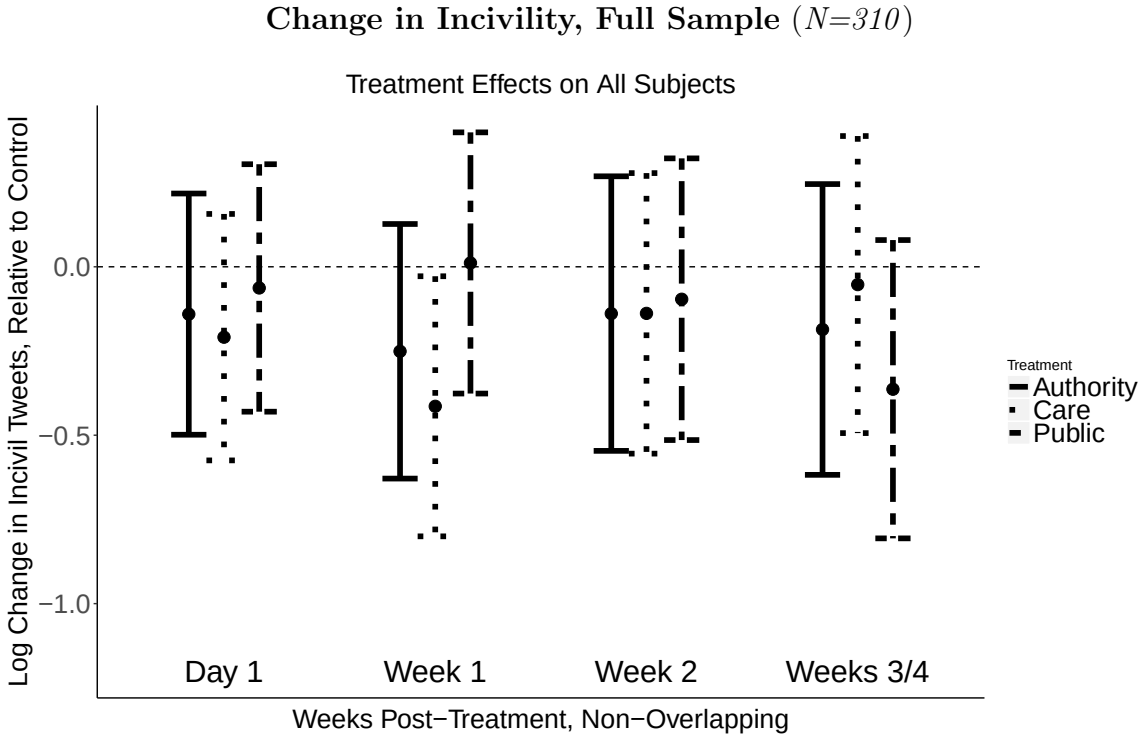


Figure 6: Treatment effects on the entire sample.

Figure 6 further disaggregates the treatment effects. In the main Week 1 time period, there is a significant effect of the Care treatment, but not the Authority treatment. These treatment effects are not significantly different from each other, however.

To test the first portion of Hypothesis 1, Figure 7 breaks down Figure 6 by the partisanship of the subjects. The top panel displays the results for Republicans, and the bottom panel for Democrats.

These results do not support the partisan portion of Hypothesis 1.

To test Hypothesis 2, I re-ran the analysis in Figure 6 with an interaction term between subject anonymity (on the three-point scale, where 0 means they provided a full bio and 2 means they were fully anonymous). Figure 8 reports the results for the Day 1 (top panel) and Week 1 (bottom panel) time periods.

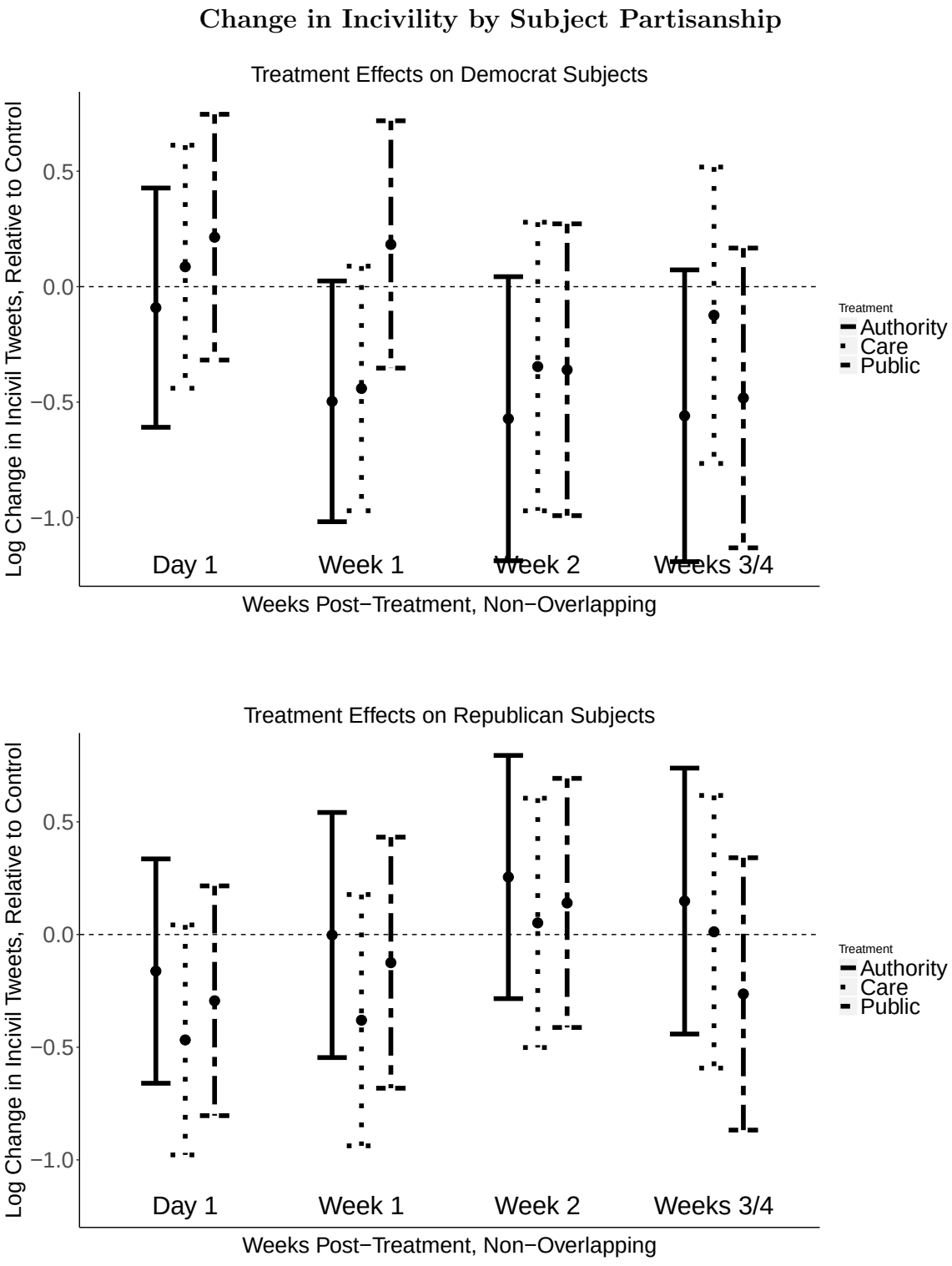
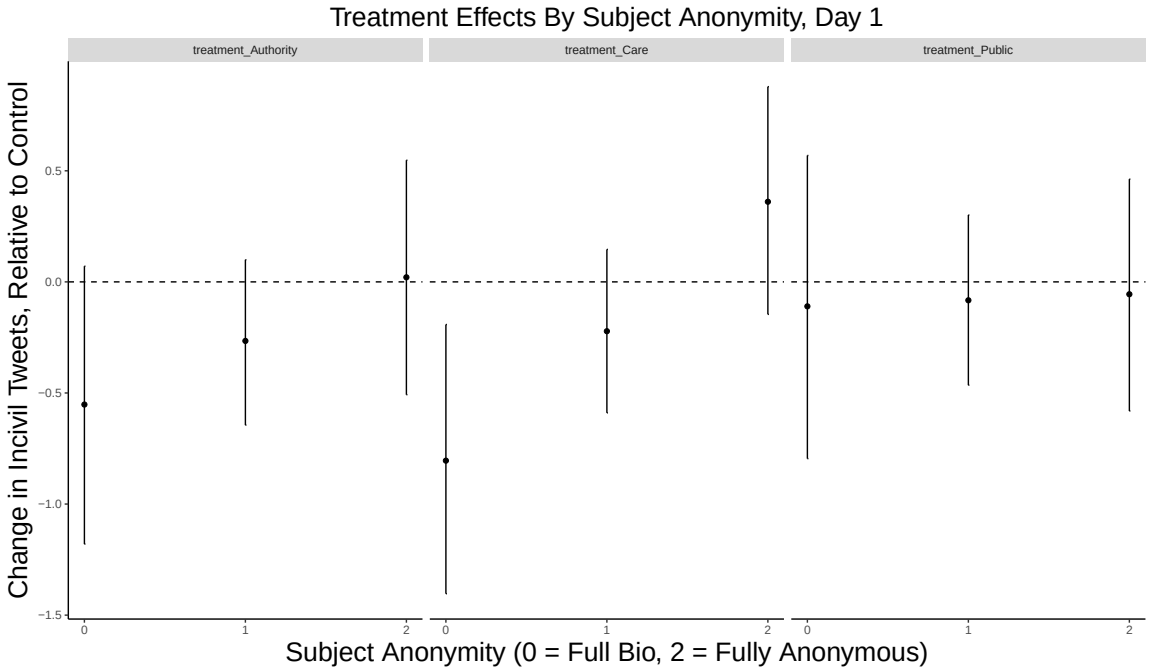


Figure 7: Treatment effects divided by subject partisanship.

Change in Incivility by Subject Anonymity (N=310)



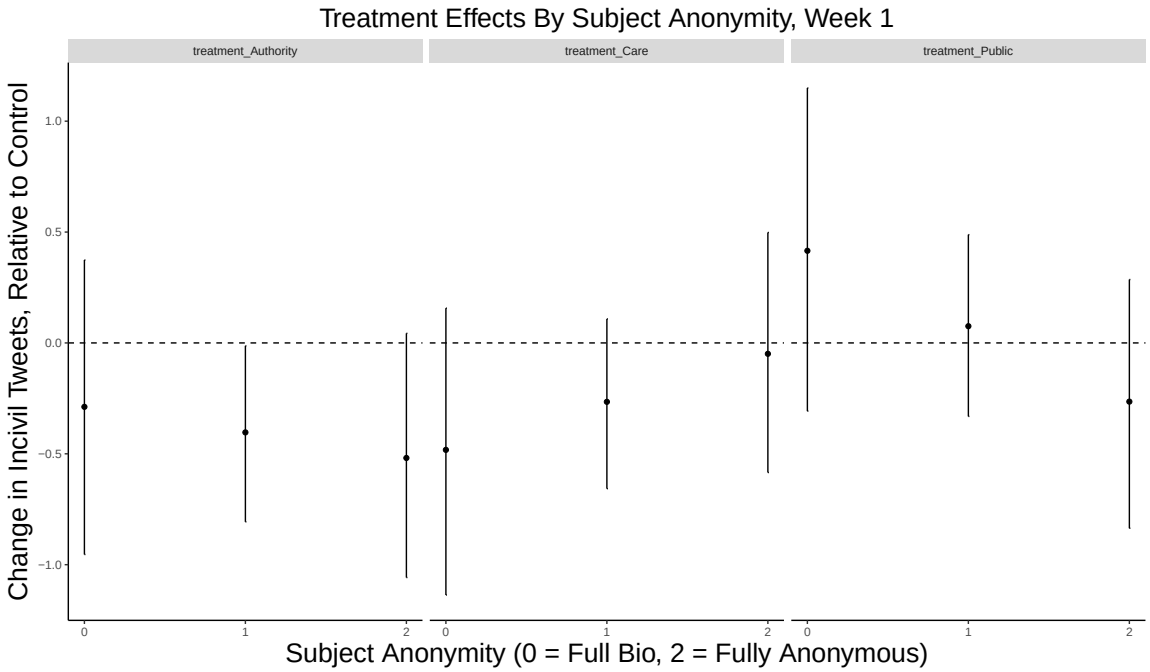


Figure 8: Treatment effects divided by subject anonymity.

In the Day 1 time period, subject anonymity behaves as expected for the two moral treatments: less anonymous subjects change their behavior more by sending fewer uncivil tweets. This interaction effect is statistically significant for the Care treatment at $p < .05$, but not for the Authority treatment ($p=.11$).

This trend does not obtain in the Week 1 time period. There is no evidence of any interaction effect for the Authority or Public treatment conditions, and the evidence for such an effect with the Care condition is very weak. There were no significant treatment effects for the later time periods, and similarly no evidence of heterogeneous treatment effects (results not shown).

One possible explanation for the lack of the expected effect on Democrat subjects is that this group was more heterogeneous. I explore this possibility in Appendix H. I find that some of the accounts coded as Democrats were actually anti-Trump Republicans, who were understandably unmoved by an appeal to the Democrat in-group.

3 Conclusion

The 2016 US presidential election took place during a time of rapidly changing norms of political civility. Although civility had been associated with conservatism, especially during the latter half of the 20th century, the campaign of Donald Trump embraced incivility while Hillary Clinton explicitly positioned herself as the candidate of civility.

The idiosyncrasies of this election may explain my failure to find support for the first half of Hypothesis 1: there was little difference between sanctioning language designed to appeal to Democrats' and Republicans' moral sense of Care or Authority. The small sample size of the experiment does not allow me to rule out the existence of moderate differences in treatment effects by partisanship, but I do find evidence against massively effective interventions targeted at subjects' specific moral framework.

The *post hoc* explanation for why Trump supporters did find the argument that they should "behave according to the proper rules of political civility" compelling is straightforward. More subtly, the minimal response from Democrats to the Care treatment may be explained by the tweets they sent to my bots in response to being sanctioned. In several cases, Democrats told my bots something like "these other people are Trump supporters, so I don't care about their feelings"; no Republicans expressed a similar sentiment. This is in keeping with the theory developed in Haidt (2012): the morality of Democrats in the US is based largely on care for specific historically disadvantaged groups, a category which does not include Trump supporters.

However, I do find support for the second half of Hypothesis 1: overall, the two moral messages caused a significant reduction in incivility in the first week after treatment, and this estimated effect was significantly larger than that of the non-moral treatment. In short, despite the oddities of the election, my intervention successfully reduced partisan incivility.

Importantly, this reduction was not observed in the models that examine only the first day after treatment. There are two explanations for this: first, this 24-hour time period is simply too noisy because the sample of tweets is too small (see Appendix D for a discussion of this point). Second, this is the period in which individuals responded directly to the treatment message. For a subset of subjects, this entailed a short-term increase in incivility that counterbalanced the reduction in other subjects in this time period. Had my research design been restricted to this time period, my inferences would have been very different. Future research in this area should aim to measure effect persistence.

Supplemental Information: Table of Contents

A Attrition

Although I initially recorded 330 subjects as belonging to either a treatment or control condition, the final analysis includes only 310 subjects. The sample suffered from attrition from one of four sources.

In the case of four subjects, I mis-applied the treatment. When I used my bots to tweet at the subjects, I made a computer error and tweeted directly at them rather than in response to a specific uncivil tweet. I became aware of this possibility when one subject responded to my tweet in confusion; in re-checking the rest of the subjects, I found the other 3 mistakes.

I identified the rest of the potentially problematic subjects through patterns in their tweeting behavior. I manually re-inspected all of the profiles of subjects for whom I collected fewer than 50 tweets pre-treatment *and* 50 tweets post-treatment. The majority of the profiles I identified this way still merited inclusion; they were just people who did not tweet very often. However, I excluded others from the final sample. I did this manual re-inspection before calculating any of the results and without knowledge of the treatment condition to which the subjects belonged.

The most common problem was that I had 0 pre-treatment tweets for a subject despite having thousands of post-treatment tweets. This was caused by the timing of when I scraped their profiles and the Twitter API's historical tweet limit: Twitter will only give you the 3,200 most recent tweets from a given account. I performed a full scrape of each account within a week of the treatment. This implies that these accounts were tweeting thousands of times a week. This is very difficult for a human to do, so I suspect that many of these accounts were bots; if they were not bots, they were extremely atypical Twitter users. However, this was the single largest source of attrition. Just under 3% of the original accounts were excluded for this reason.

There were a total of 3 accounts in my sample that were suspended by Twitter

Table 2: Attrition Rates and Causes

	Control	Democrats	Republicans
Initial assignment	108	104	118
Failed treatment application	0	2	2
Tweeted too often/bots	3	1	5
Suspended	0	1	2
Weird	2	0	0
Final	102	100	108
Attrition	6%	4%	8%

during the course of my experiment. I do technically have enough tweets from these accounts to include them in the analysis, but doing so has the potential to bias my results upwards: the reduction in the number of uncivil tweets they sent was actually caused by Twitter preventing them from tweeting, rather than by the treatment.

Finally, there were two accounts that were just weird; they had not tweeted thousands of times, but each still only recorded 3 pre-treatment tweets. In both cases, the accounts appeared to be behaving very oddly, and since I did not have a reasonable estimate of their pre-treatment behavior, I excluded them.

B Empirical distribution of aggression scores in subject tweets

As shown in Figure 9, the distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain, and Dixon (2017)) is bimodal: there is a large cluster of "non-aggressive" tweets near 0, and a smaller cluster of "definitely aggressive" tweets near 1. The vertical line represents the 70th percentile of this empirical distribution, the cutoff I use in the body of the paper for transforming these scores into a binary measure. The higher cutoff of the 90th percentile would entail including only the far-

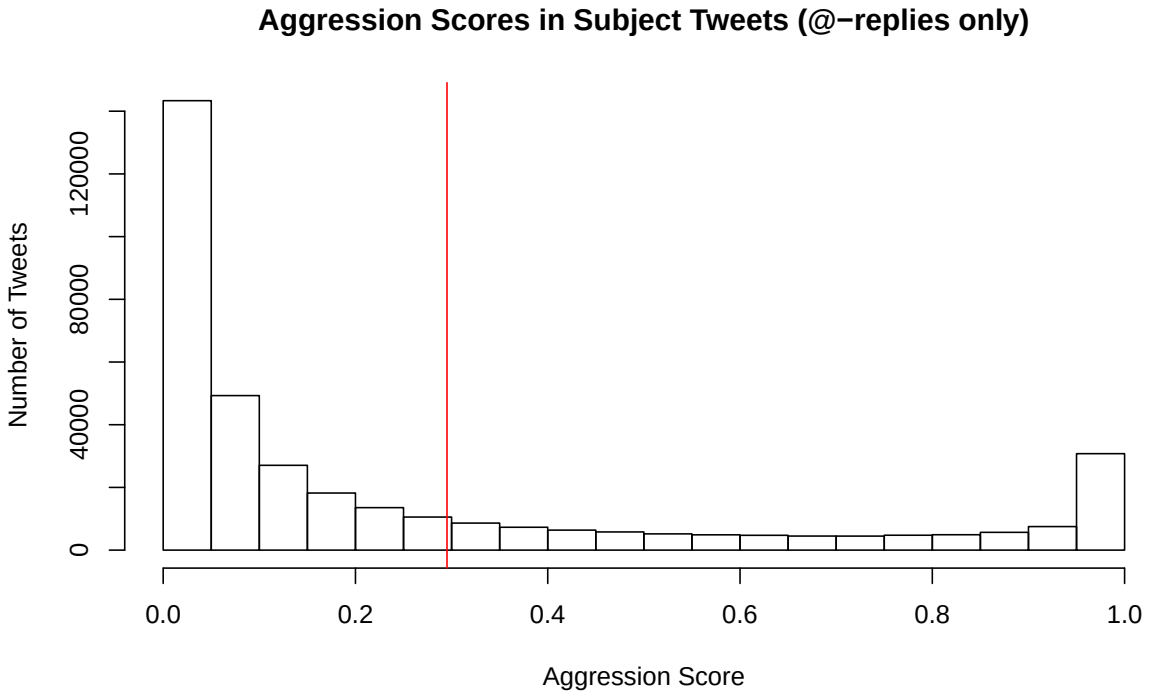


Figure 9: Empirical distribution of aggression scores. The vertical line represents the 70th percentile, the cutoff I use in the body of the paper.

right cluster of tweets. The main results are replicated using this higher threshold in Appendix D.

C Validation of Wikipedia measure on the current dataset

Figure 10 plots the accuracy of the scores derived from the Wulczyn, Thain, and Dixon (2017) model in predicting the labels of tweets coded by crowdworkers. The x-axis plots the threshold used to turn the continuous scores output by the model into binary labels. There is a slight peak (accuracy = .82) at the black vertical line, which depicts the 75th percentile, but the accuracy is fairly constant across a wide range of cutoffs.

The validation tweets consist of 1,000 tweets which were randomly sampled from among all subject tweets and uploaded to Mechanical Turk. Each tweet was coded by two of Amazon's "Expert Coders," a restrictive label that they only award to consistently attentive crowdworkers. The precise instructions given to the workers were as follows:

```
Please read each tweet and tell us if it is civil or incivil.

We say that "civil" tweets are those that demonstrate respect for the
person being tweeted at.

If a tweet has very little information (if it just contains a link, for
example), code it as "civil."
```

Overall levels of intercoder reliability were low by the standards of objective classification tasks (Krippendorf's alpha = .37). The task at hand, however, is inherently subjective, and our results are in line relevant published work: Wulczyn, Thain, and Dixon (2017), using a somewhat more rigorous coder vetting process, report "a Krippendorf's alpha score of 0.45. This result is in-line with results achieved in other crowdsourced studies of toxic behavior in online communities. (p3)"

For the accuracy results displayed in Figure 10, I restricted the initial 1,000 tweets to the 70% on which the coders agreed on the label. These labels are unbalanced in the sample (74% were labeled civil), so the 82% accuracy represents a significant improvement on a naive classification scheme.

As the confusion matrices below indicate, maximizing accuracy entails a tradeoff with balancing classification discrepancies. At the 70th percentile threshold, the percentage of validation tweets labeled as uncivil by the human coders but civil by the

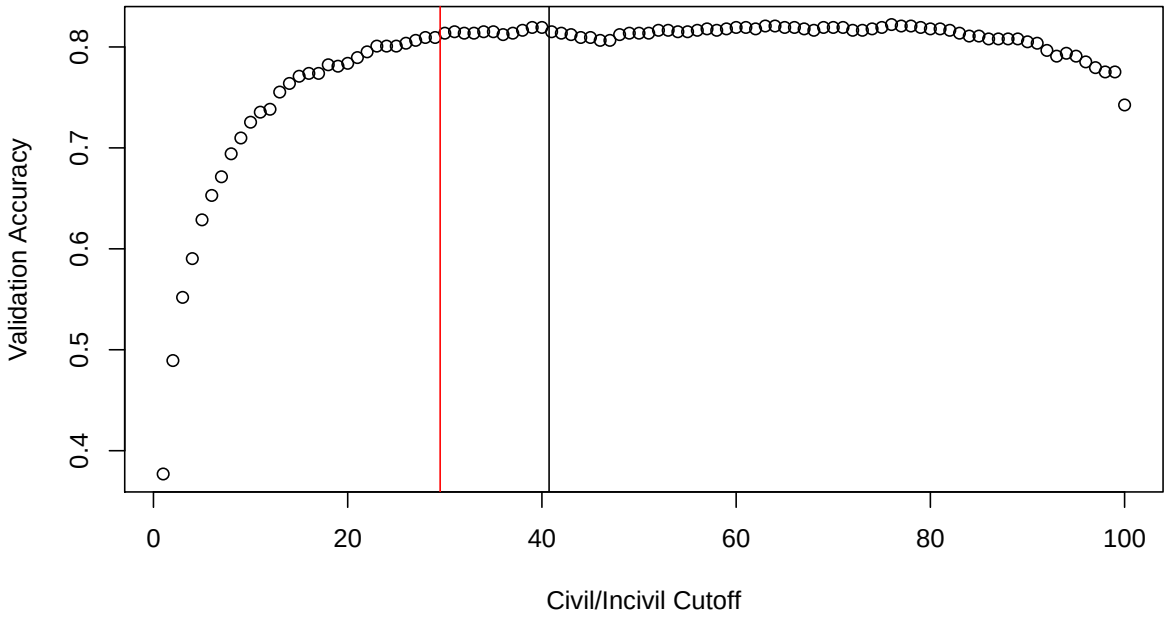


Figure 10: Accuracy of the Wikipedia model applied to tweets labeled by Mechanical Turk workers, scored on the tweets on which coders agreed on whether the tweet should be labeled civil or incivil. The red vertical line represents the 70th percentile, the cutoff I use in the body of the paper.

Confusion matrix of human and algorithmic labels on validation tweets; 75th percentile

	Mturk: Civil	Mturk: Incivil
Algorithm: Civil	67.7%	11.8%
Algorithm: Incivil	6.5%	13.9%

Confusion matrix of human and algorithmic labels on validation tweets; 70th percentile

	Mturk: Civil	Mturk: Incivil
Algorithm: Civil	65.3%	9.8%
Algorithm: Incivil	9.0%	15.9%

algorithm is 11.8%, compared to 6.5% for disagreements in the opposite direction.

To balance these discrepancies, the second confusion matrix reports the results when the incivility threshold is lowered to the 70th percentile. This balancing comes at the cost of lowering the overall accuracy from 81.8 to 81.2. On balance, I believe that this slight decrease in accuracy is less important than using a measure that concords with the human coding as much as possible, so I use the 70th percentile threshold in the body of the text.

D Main Results Using Higher Aggression Threshold

The results in the body of the paper use the 70th percentile of the empirical distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain, and Dixon (2017)) to code subject tweets as civil or incivil. There is a distinct cluster of "definitely aggressive" tweets near the top of this distribution, and the results in Figure 11 plot the model results when only this cluster is coded as incivil—that is, using the 90th percentile as the threshold. In both plots, the effects in the 1 Day time period become more pronounced to 0, while the effects in the 1 Week time period become closer to 0 and not statistically significant.

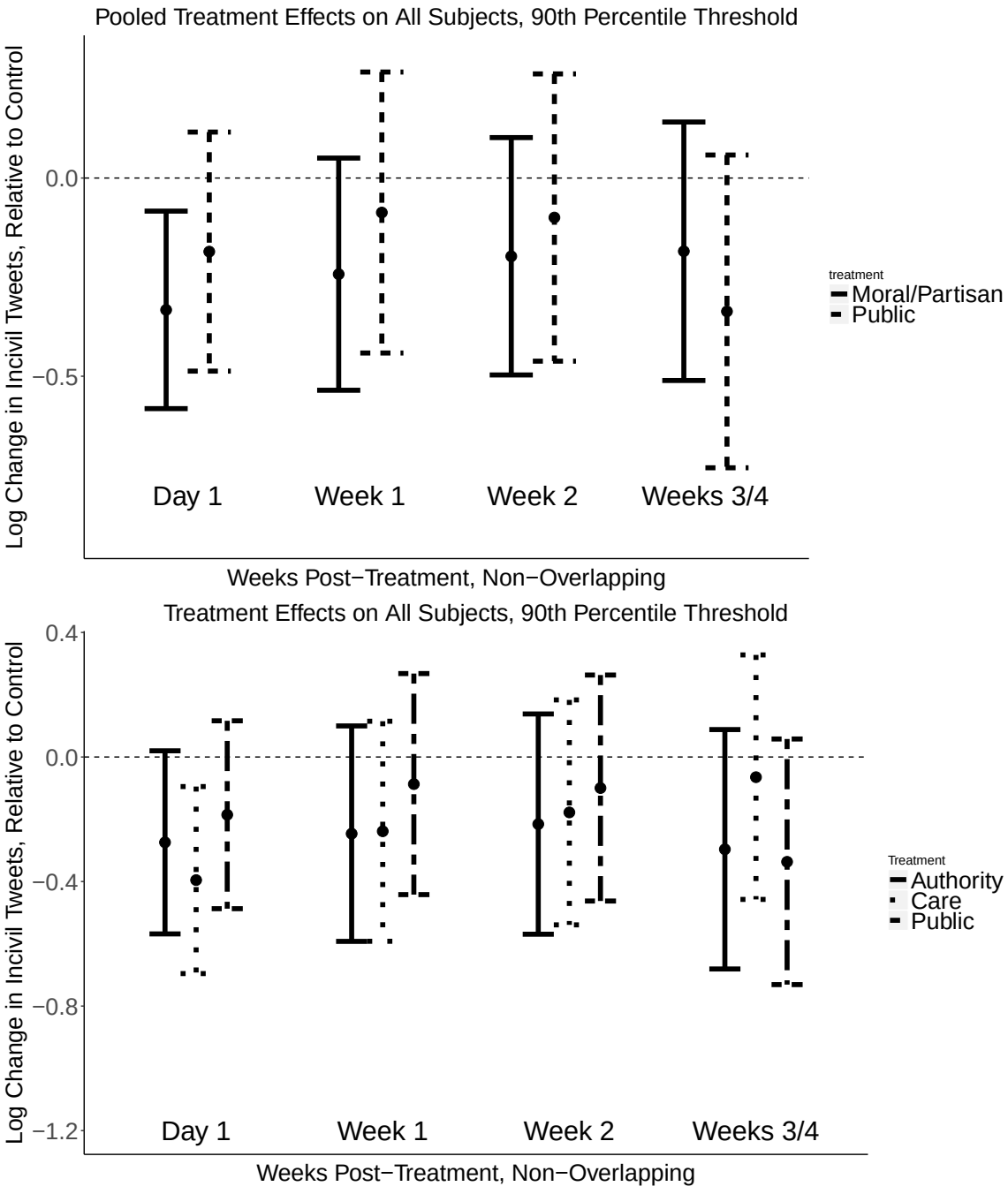


Figure 11: Main results replicated using the higher threshold of aggression scores for coding tweets as incivil.

E Negative Binomial Specification of Main Results

The dependent variable of interest in this analysis is the number of times a subject sent an uncivil tweet to another user. This is a "count variable"–it can only take non-negative integer values–and thus violates a fundamental assumption of OLS regression. To address this issue, generalized linear models with different assumptions are often used. Poisson regression, in which the dependent variable is assumed to have a Poisson distribution, is a common technique, but this carries the further assumption that the variance and expected value of the dependent variable are equal. In cases in which the variance is significantly higher than the expected value–like it is here–the negative binomial model relaxes this assumption (Hilbe, 2011).

$$ln(Agg_{post}) = x_{int} + \beta_1 Agg_{pre} + \beta_2 T_{feel} + \beta_3 T_{rules} + \beta_4 T_{public} + \beta_5 Anon + \beta_6 (T_{feel} \times Anon)$$

$$+\beta_7 (T_{rules} \times Anon) + \beta_8 (T_{public} \times Anon)$$

To interpret the relevant treatment effects implied by the coefficients estimated by this model, the exponent of the estimated $\hat{\beta}_k$ for each of the treatment conditions needs to be added to the corresponding $\hat{\beta}$ for the interaction term, evaluated at each level of Anonymity Score (Hilbe, 2011). For example, the effect of the Feelings treatment on subjects with Anonymity Score 1 (the middle category) is:

$$IRR_{feel \times Anon_1} = e^{\hat{\beta}_2 + \hat{\beta}_6 \times 1}$$

The results of these negative binomial models can be seen in Figure 12 and Figure 13.

Change in Incivility, Full Sample, Negative Binomial Specification (*N=310*)

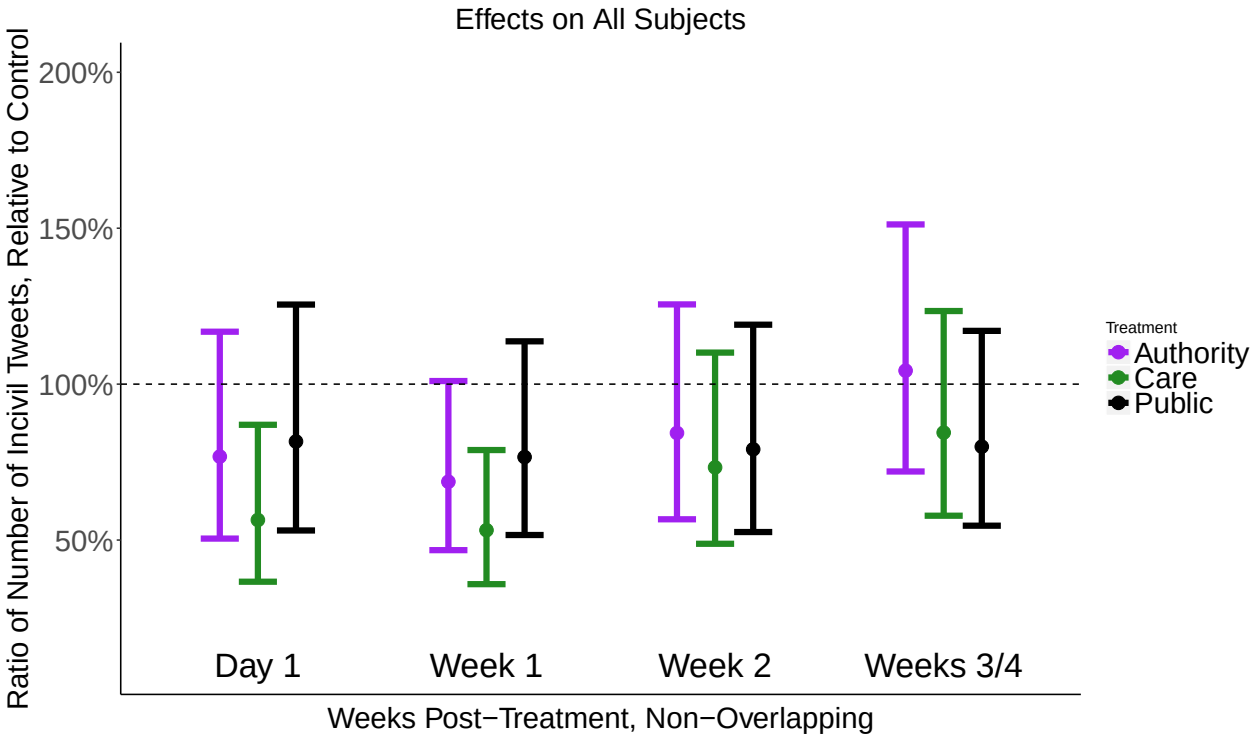
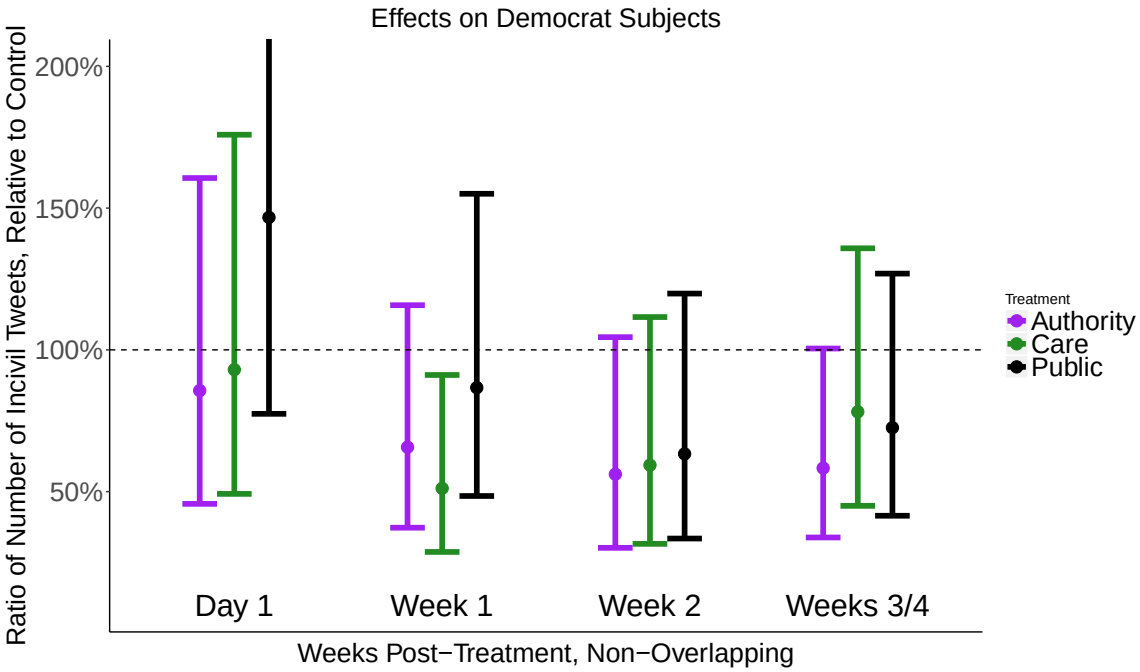
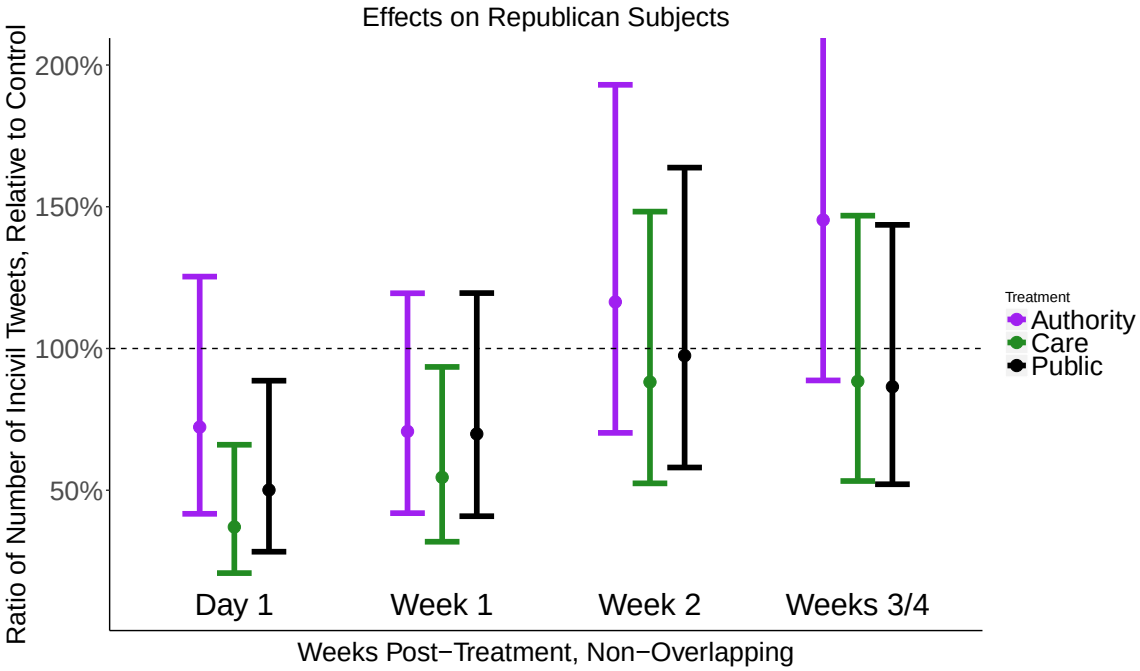


Figure 12: The Incidence Ratio calculated from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 50% as many directed uncivil tweets as the subjects in the control group. 95% confidence intervals.

Figure 13: **Effects on Democrats, Negative Binomial Specification** *(N=147)*



Effects on Republicans, Negative Binomial Specification *(N=163)*



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot in Panel A means that these subjects sent 90% as many directed uncivil tweets as the subjects in the control group. 95% confidence intervals.

F Results Divided by Subject Loquacity

The subject population is highly varied in their pre-treatment level of tweeting activity. Although not one of the tests I specified in my pre-analysis plan, the potential policy implications of heterogeneous effects based on subject loquacity merit investigation of this possibility.

Figure 14 replicates the main results in the paper by the pre-treatment tweeting rate of the subjects. The top panel displays the results for subjects above the median (82 uncivil pre-treatment tweets), and the bottom panel for subjects blow this threshold.

There is a clear distinction: treatment effects on the more active subjects are close to zero, while the effects (of the moral treatments) on the less active subjects are significant in the 1 week time period.

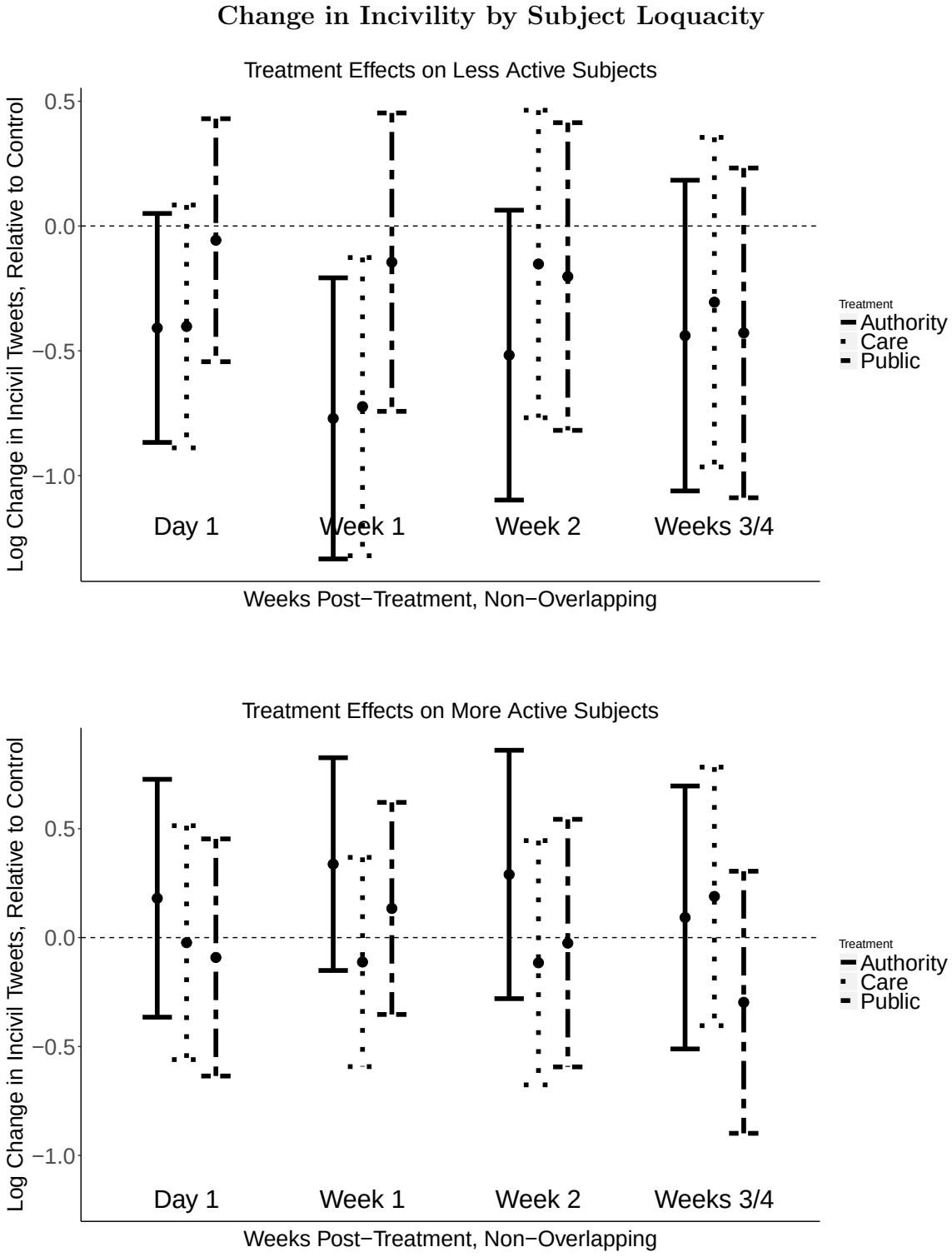


Figure 14: Treatment effects divided by subject pre-treatment tweeting rate. The top panel displays results for less active subjects (below the median), while the bottom panel displays results for more active subjects (above the median). Lines represent 95% confidence intervals.

G Treatment Effects on Sending Civil Tweets

The results in the body of the paper display treatment effects on the rate of sending *incivil* tweets. It is worth exploring whether the treatment had an analogous effect on sending *civil* tweets.

I re-ran the analysis using the number of *civil* tweets as the dependent variable (those with aggression scores below the 70th percentile threshold), and found no significant treatment effects. The point estimates are in the same direction as the effects on uncivil tweets, with effect sizes ranging from 50% to 80% as large. Figure 15 displays these results. In Panel A, examining pooled treatment effects, these effect sizes are for the civil tweets, .08 (1 Day) and .16 (1 Week); for the uncivil tweets in the body of the text, these effect sizes are, .15 and .2.

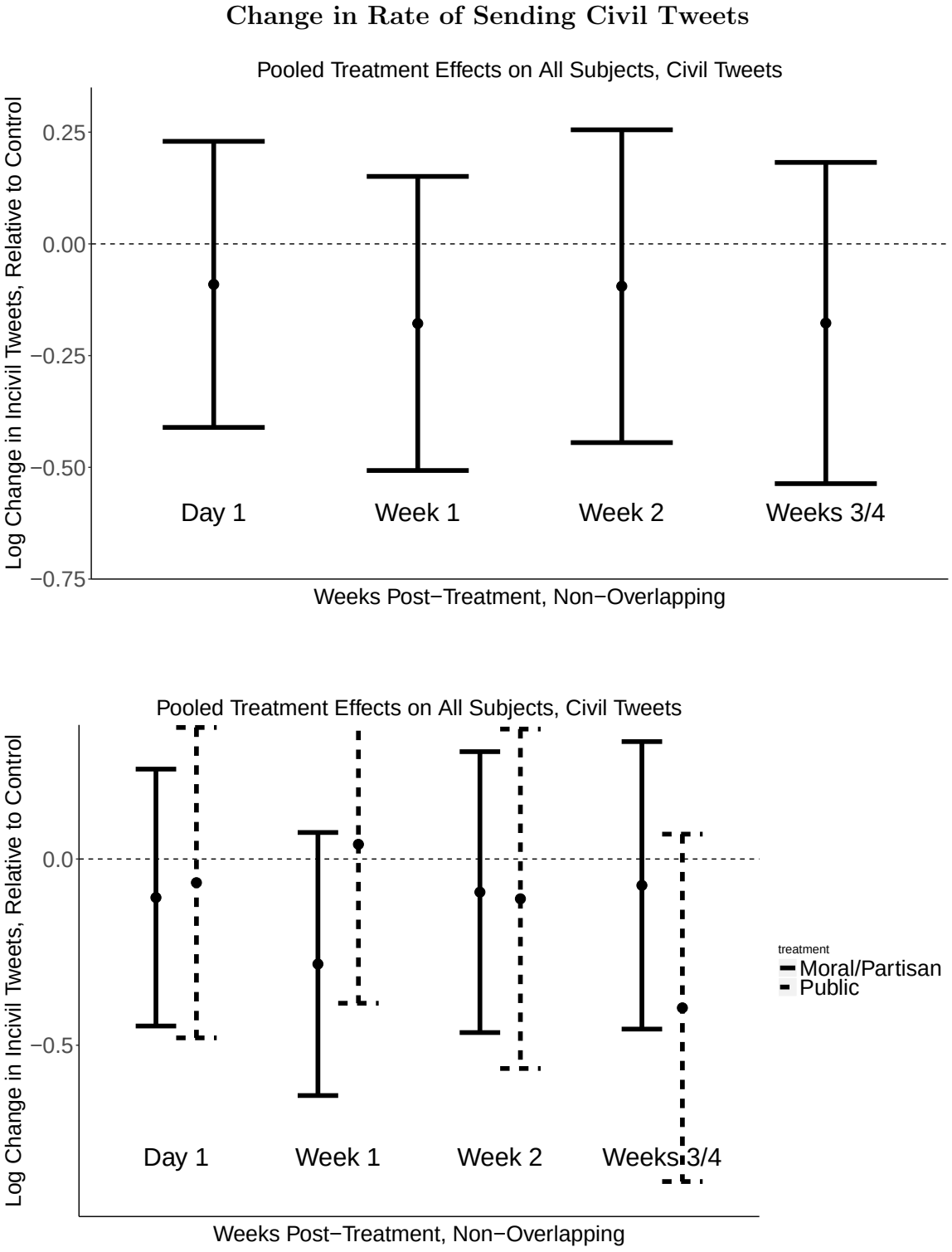
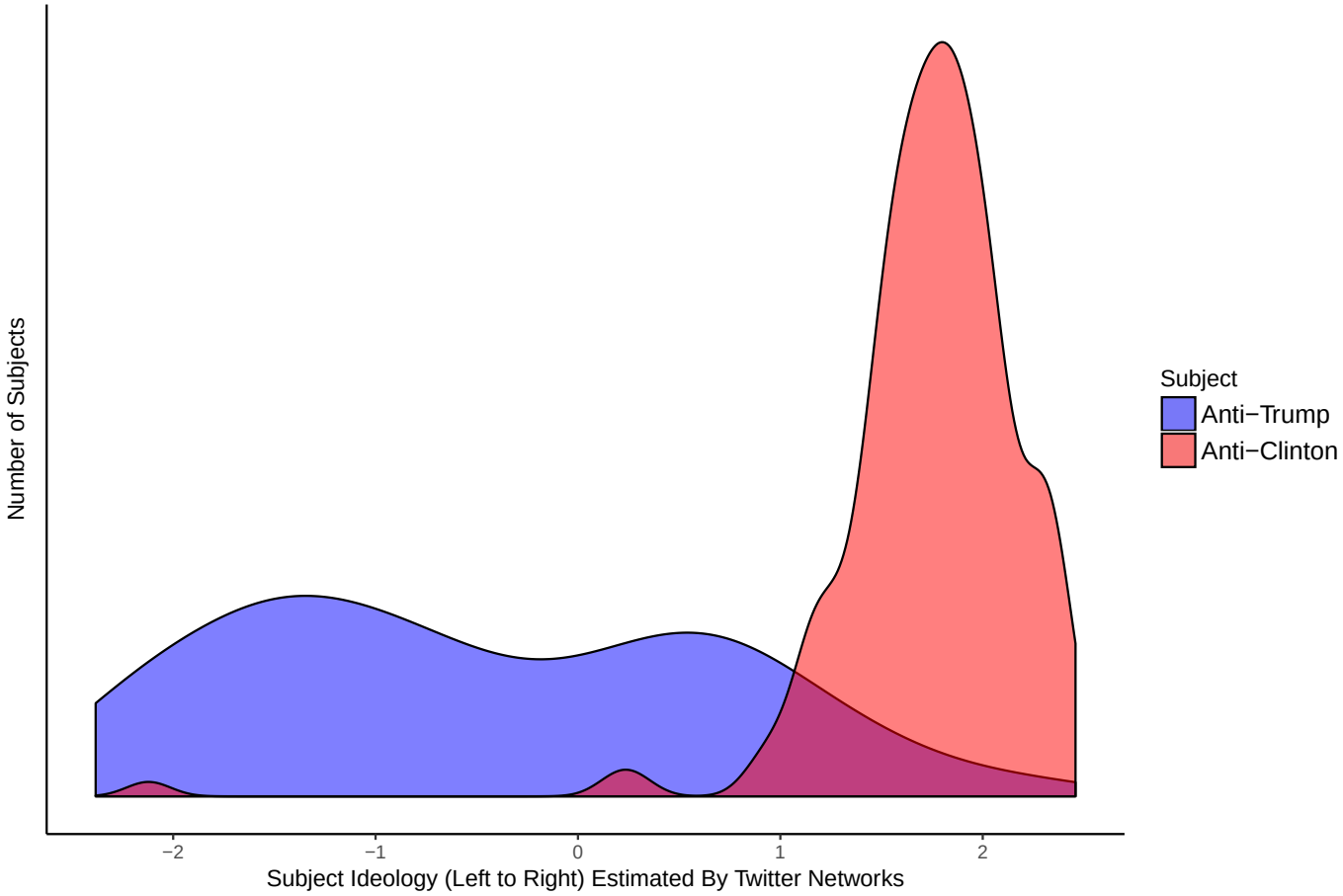


Figure 15: Treatment effects on the rate of subjects sending civil tweets (tweets scored as below the threshold for incivility used the body of the paper). The top panel displays results pooled across all treatment conditions, while the bottom panel displays results where the two moral treatments are pooled. Lines represent 95% confidence intervals.

Figure 16: Estimated Ideology of Subjects Labeled "Republican" or "Democrat"



H Ideological Heterogeneity

I implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. As Figure 16 demonstrates, there was significant heterogeneity in the ideal points of subjects I coded as Democrats, but not for Republicans.

All but two of the subjects coded as Anti-Hillary (Republicans) had estimated ideology scores above 1. However, a full third of the subjects coded as Anti-Trump (Democrats) had estimated ideology scores right of center, although only a few are far to the right (have an ideology score above 1). Looking at Figure 16, there appears to be two distinct clusters of Anti-Trump subjects. In addition to the expected group of

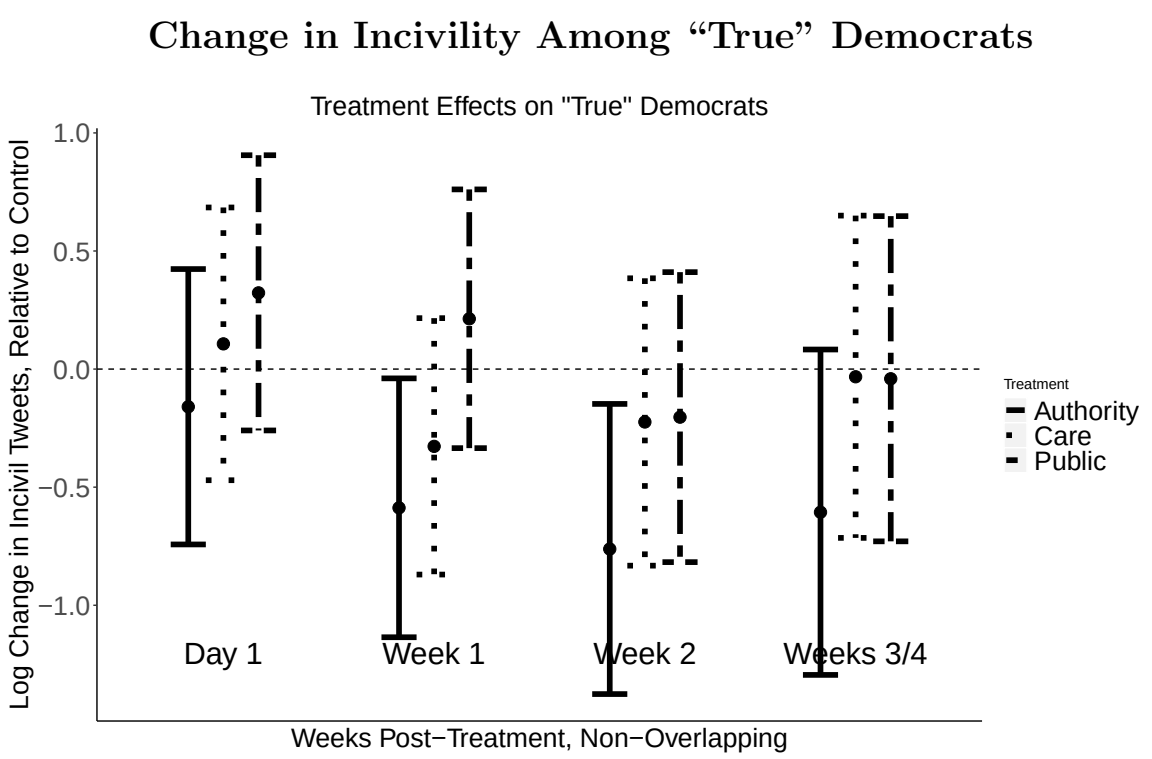


Figure 17: Treatment effects on Democrat subjects, restricted to subjects whose ideologies were estimated to be left of the anti-Clinton cluster on the right (*N=122*).

Democrats, there is also a significant contingent of moderate Anti-Trump Republicans that I classified as Democrats. Because the Care and Authority treatment messages were explicitly designed to appeal to subjects' partisan group identities (and identified the Anti-Trump subjects as "Democrats"), the ideological heterogeneity within this group could pose a problem for estimating average treatment effects.

If I restrict the analysis of Democrats in Figure 7 to only those with estimated ideology scores to the left of the major cluster of anti-Clinton subjects in Figure 16, I find some support for this *ex post* explanation. The point estimates for the Authority treatment effect becomes more negative in the Week 1 and Week 2 time periods, seen in Figure 17. Because the sample size is down to 86, the Care treatment is still not significant, but the largest change is on the Authority effects, which are now significantly negative in the Week 1 and Week 2 time periods.

References

Barberá, Pablo. 2014. "streamR: Access to Twitter Streaming API via R." *R package version 0.2* 1.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91.

Bejan, Teresa M. 2017. *Mere Civility*. Harvard University Press.

Berry, Jeffrey M, and Sarah Sobieraj. 2013. *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Chen, Adrian. 2015. "The Agency." *New York Times Magazine* June 2, 2015.
URL: *https://www.nytimes.com/2015/06/07/magazine/the-agency.html*

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee pp. 963–972.

Duggan, M, and A Smith. 2016. "The political environment on social media." *Pew Research Center* .
URL: *http://www.pewinternet.org/2016/10/25/political-content-on-social-media/*

Frijda, Nico H. 1988. "The laws of emotion." *American psychologist* 43 (5): 349.

Greenwood, S, A Perrin, and M Duggan. 2016. "Social Media Update 2016." *Washington, DC: Pew Internet & American Life Project. Retrieved November* 27: 2016.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108 (4): 814.

Haidt, Jonathan. 2012. "The Righteous Mind: Why good people are divided by religion and politics." *Pantheon, New York* .

Hilbe, Joseph M. 2011. *Negative binomial regression*. Cambridge University Press.

Hindman, Matthew. 2008. *The myth of digital democracy*. Princeton University Press.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, not ideology a social identity perspective on polarization." *Public opinion quarterly* 76 (3): 405–431.

Muddiman, Ashley. 2017. "Personal and public levels of political incivility." *International Journal of Communication* 11: 21.

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39 (3): 629–649.

Munger, Kevin. 2020. "Replication Data for: Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter.".
URL: *https://doi.org/10.7910/DVN/OUYTUP*

Mutz, Diana C. 2015. *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Phillips, Whitney. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.

Theocharis, Yannis, Pablo Barberá, Zoltan Fazekas, and Sebastian Adrian Popa. 2015. "A Bad Workman Blames His Tweets? The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates." *The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates (September 5, 2015)* .

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 1391–1399.