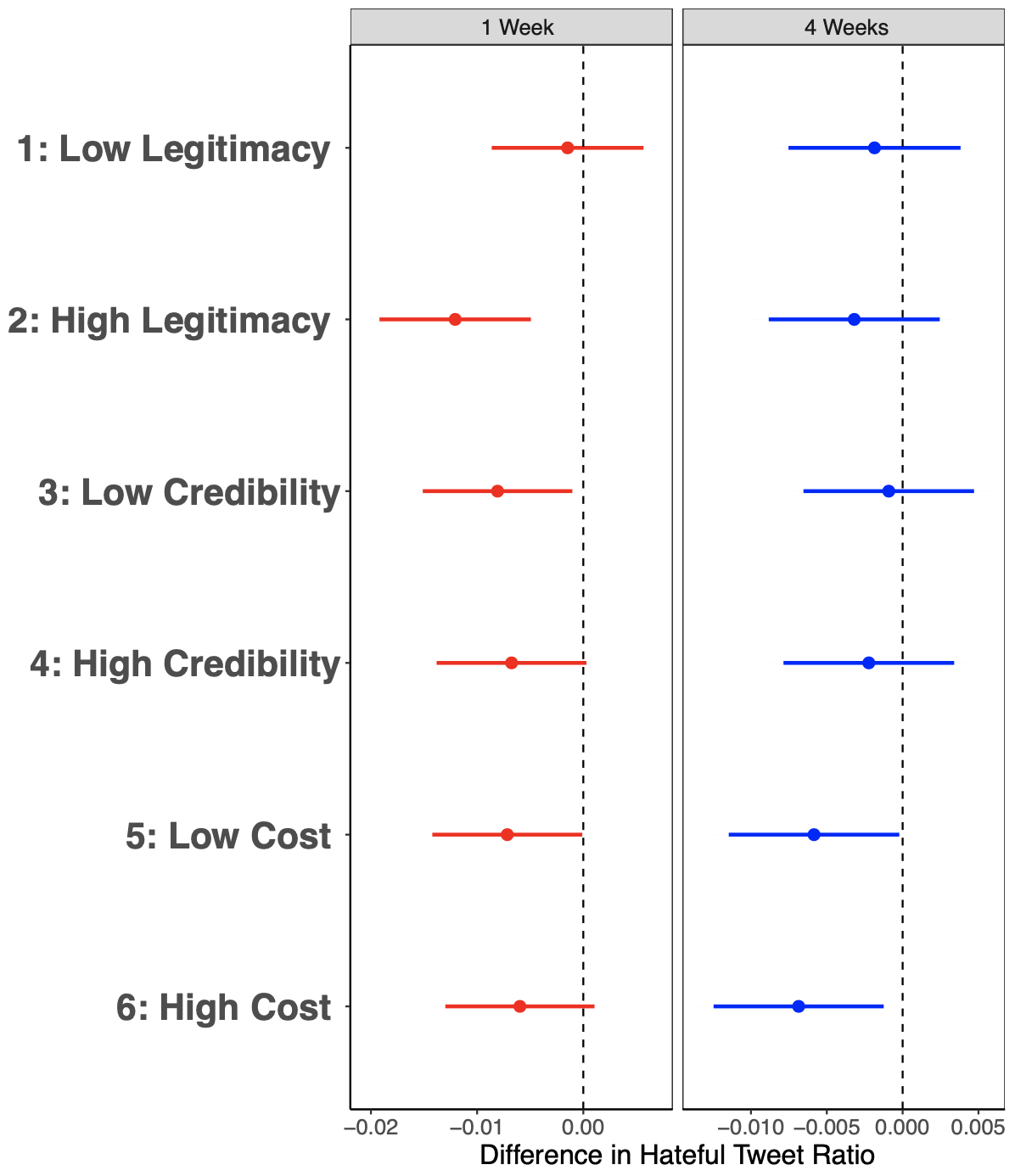
**Appendix**

**Appendix A: Missing Observations**

In our Pre-Analysis Plan (PAP), we state that we will drop the missing observations that are due to account deletion, profiles that became non-public after the treatment, or account deletion. However, as a robustness check, we show the results from imputing values for the missing observations. We follow Munger (2017) and treat users who disappeared from our sample as not having used any hateful language after the treatment. Figure A2 shows the coefficient plot from imputing the missing values with 0.



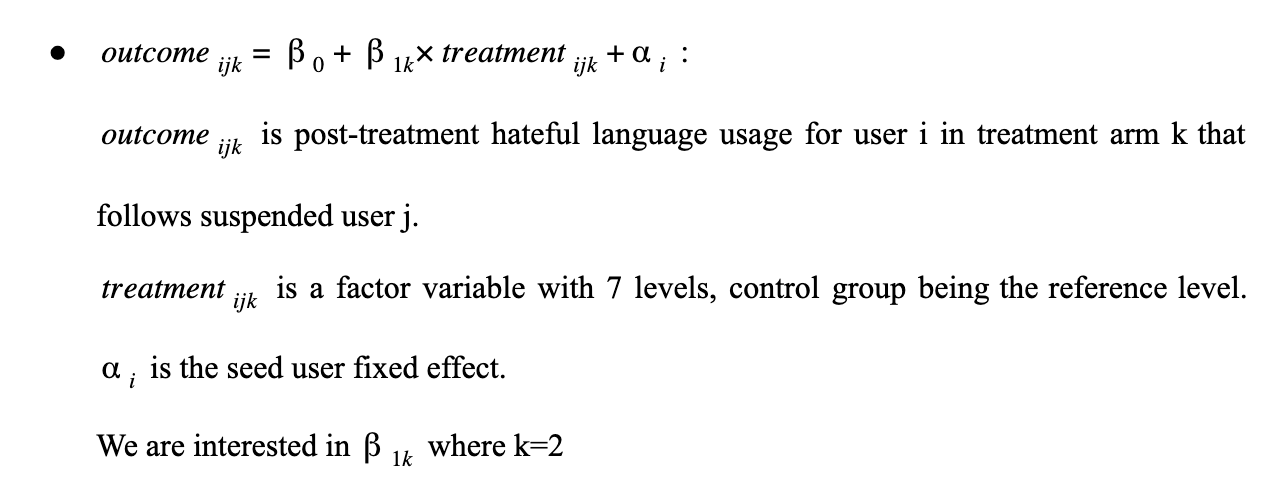
**Figure A1**

**Appendix B: Power analysis**

In order to determine the smallest sample size that is suitable to detect the effect of our treatments, we conduct power analyses. Power is the probability of rejecting the null hypothesis when it is in fact false. To be adequately powered, the power of a test should be above some prespecified level. In our power analyses, we use the conventional threshold of 80 percent (Kraemer and Blasey 2015). In what follows, we explain the process of how we obtained the sample size that would give us enough power to detect our treatment effects.

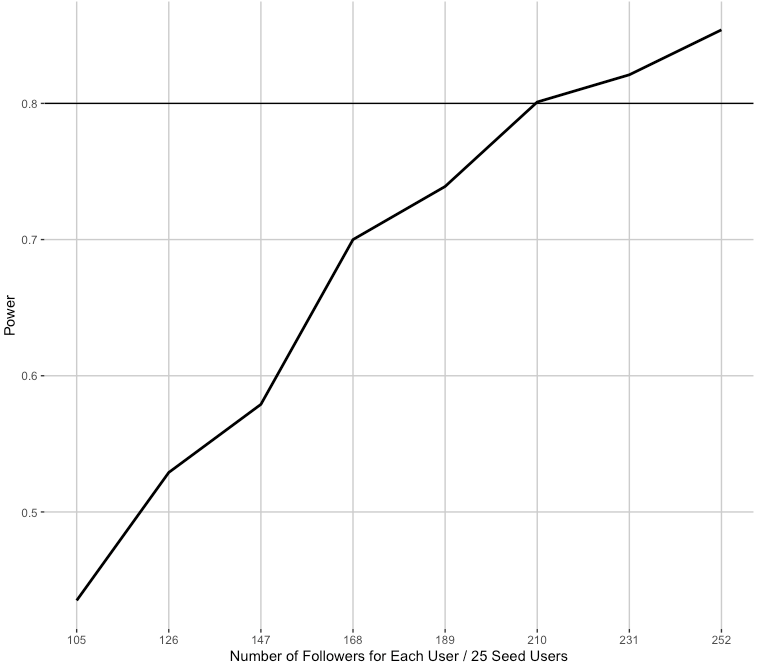
We take our hypothesized treatment effect as -0.34 for the second, fourth and sixth treatment arms. The number comes from the main finding of Munger (2017) study.

In each simulation, we create a dataset with suspended users, their followers, their assigned treatment arms, their baseline usage of hate words (normal distribution with mean 3 and sd 1 separately generated for the followers of each seed user) and simulate treatment effects (mean 0 and sd 1 for the control group, mean 0.34 and sd 1 for the treatment groups in 2nd, 4th and 6th treatment arms) for each treatment arm. After generating the data as such, we run the following specification:



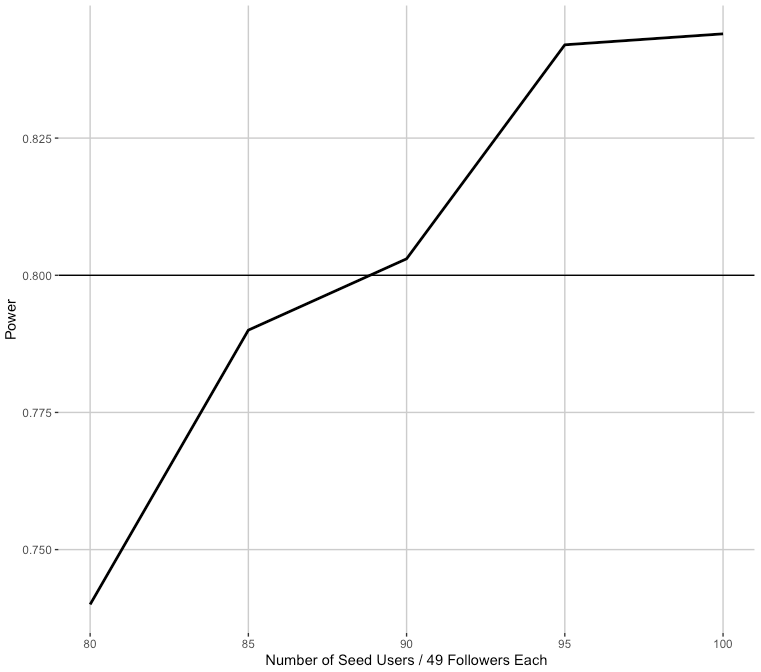
After running the specification, we record whether the p-value for B\_1\_k is under 0.05 or not. We run 1000 simulations as such and record the ratio of p-values that is under the value 0.05, which we call as power in the y-axis.

In each of the following plots, we change one parameter while keeping the others fixed. In the first plot in Figure A3, we keep the number of seed users fixed, and change the number of treated followers each seed user has. For each value, we run the simulation 1000 times and record the power levels.



**Figure B1**

In the second plot in figure A4 we keep the number of treated followers each seed user has fixed, and change the number of seed users. For each value, we run the simulation 1000 times and record the power levels.



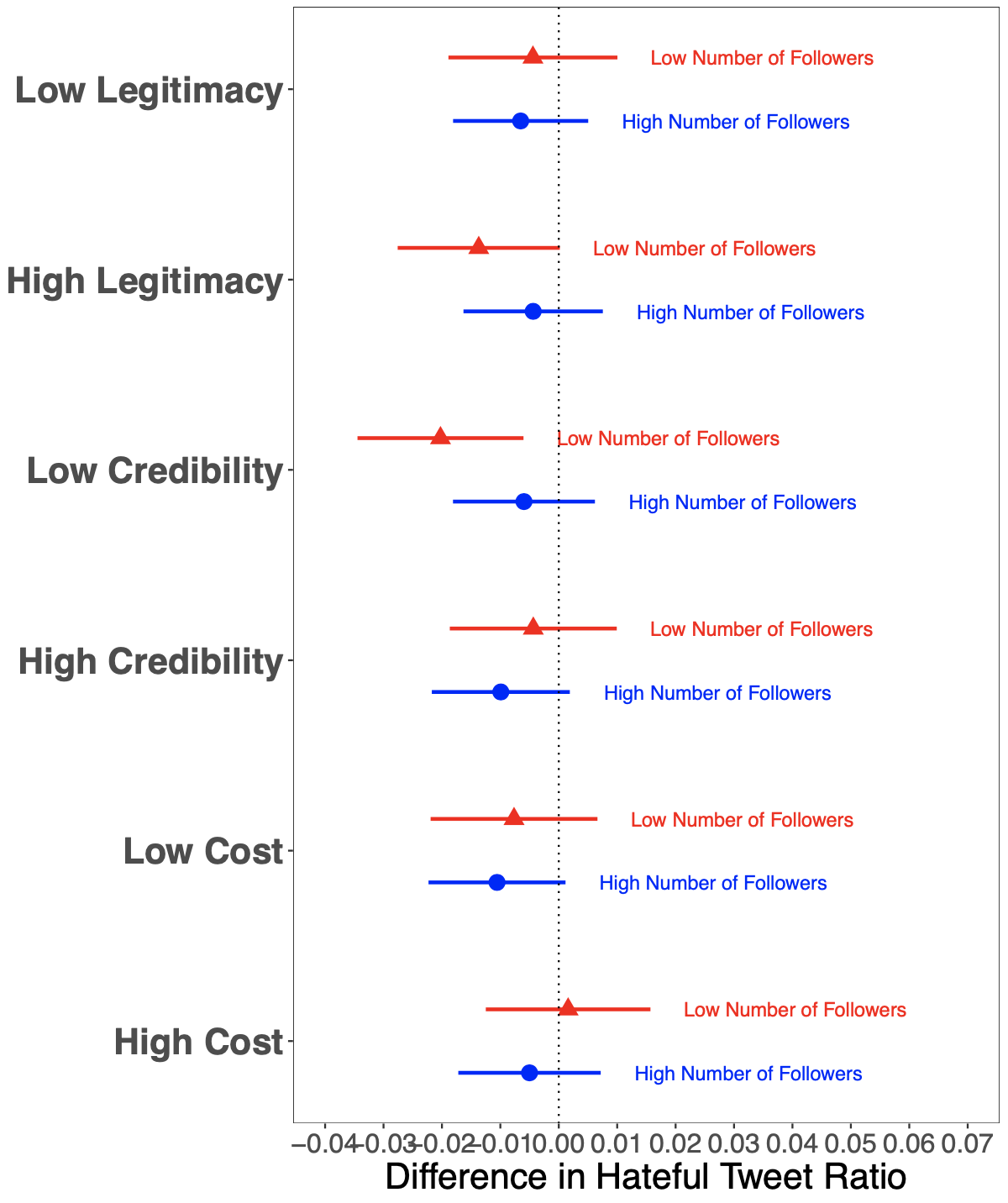
**Figure B2**

The plots show that we are statistically powered enough to detect effects at the 0.05 significance level if we have 25 seed users with 210 followers each or 87 seed users with 49 followers each. This means that with 27 seed users and in total 4510 followers, we are powered enough to detect our treatment effects.

**Appendix C: Heterogeneous Treatment Effects**

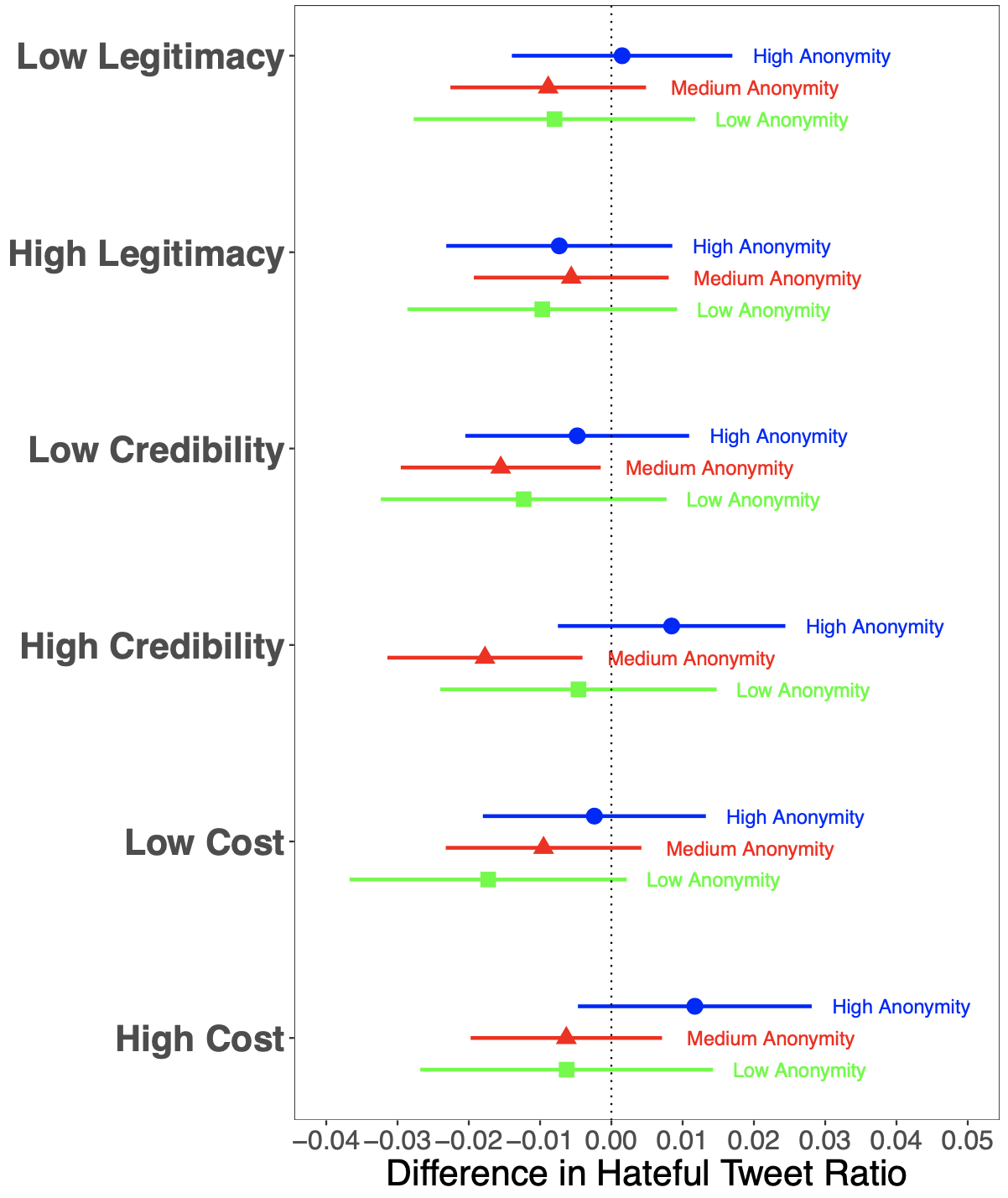
As we specify in Footnote 5, in the interest of space and at the advice of the editors, we do not present the full results regarding heterogeneous treatment effects in the body of the paper. In this section, we present heterogeneous treatment effects based on the number of followers each account has, the anonymity of the account, the activeness of the account on Twitter, and the amount of time since the creation of the account.

The coefficient plot in Figure C1 shows how the effects are moderated by the number of followers a user has. We divided our sample into two groups: users with above the median number of followers and below the median number of followers.[[1]](#footnote-0) The results do not provide support for HE1 where we hypothesized that users with higher number of followers have more to lose from a potential, and therefore would perceive their potential suspension as more costly compared to other users, which would make them more likely to reduce their hateful language after receiving our warning tweet.



**Figure C1[[2]](#footnote-1)**

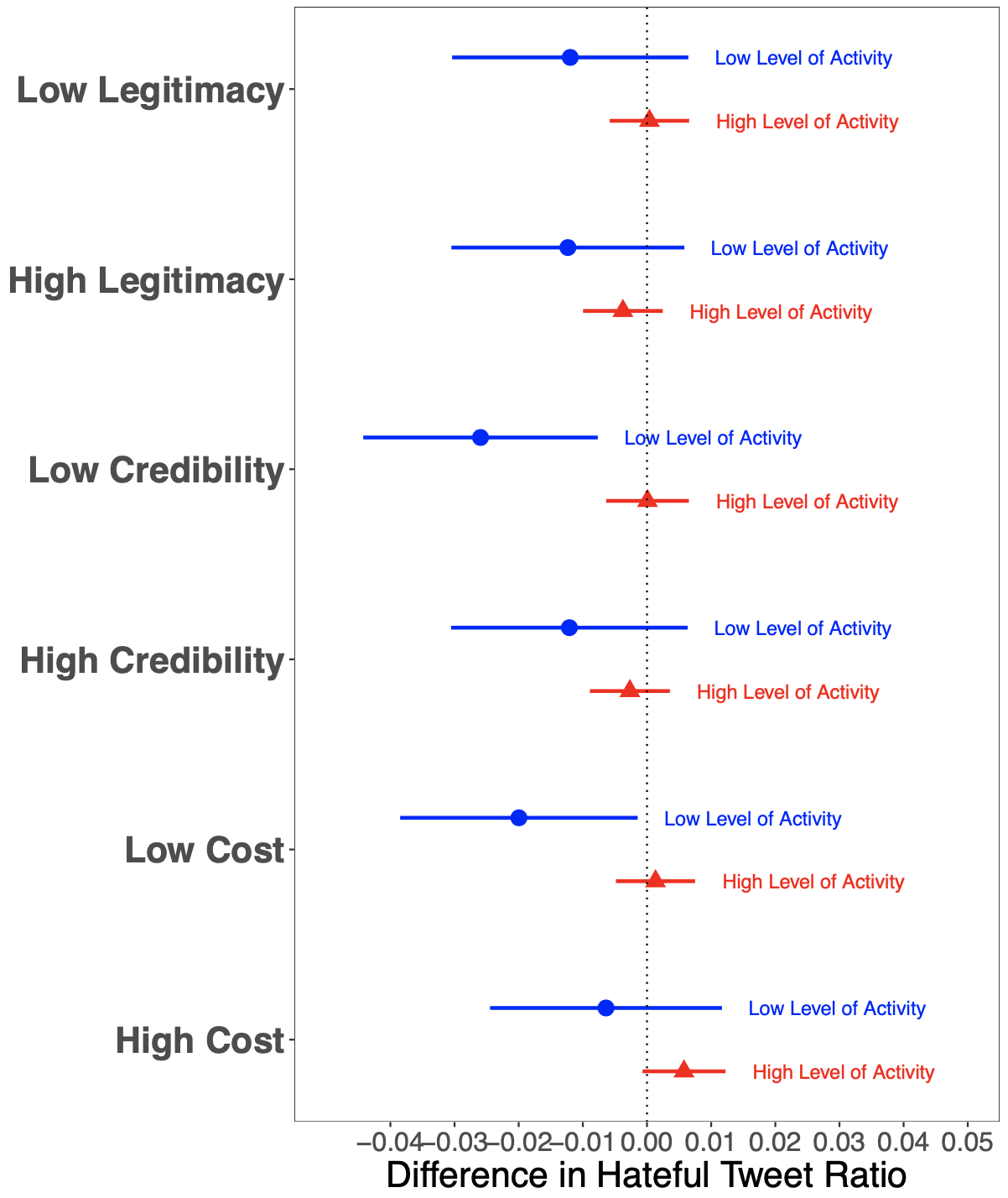
The coefficient plot in Figure C2 shows the effect of each treatment tweet on the use of tweets with hateful language depending on the anonymity level of the user. The results do not provide any support for HE2 where we hypothesized that users with less anonymous profiles would be more likely to think that the online consequences that we warn them of could also affect their offline life, and therefore would decrease their use of hateful language.



**Figure C2[[3]](#footnote-2)**

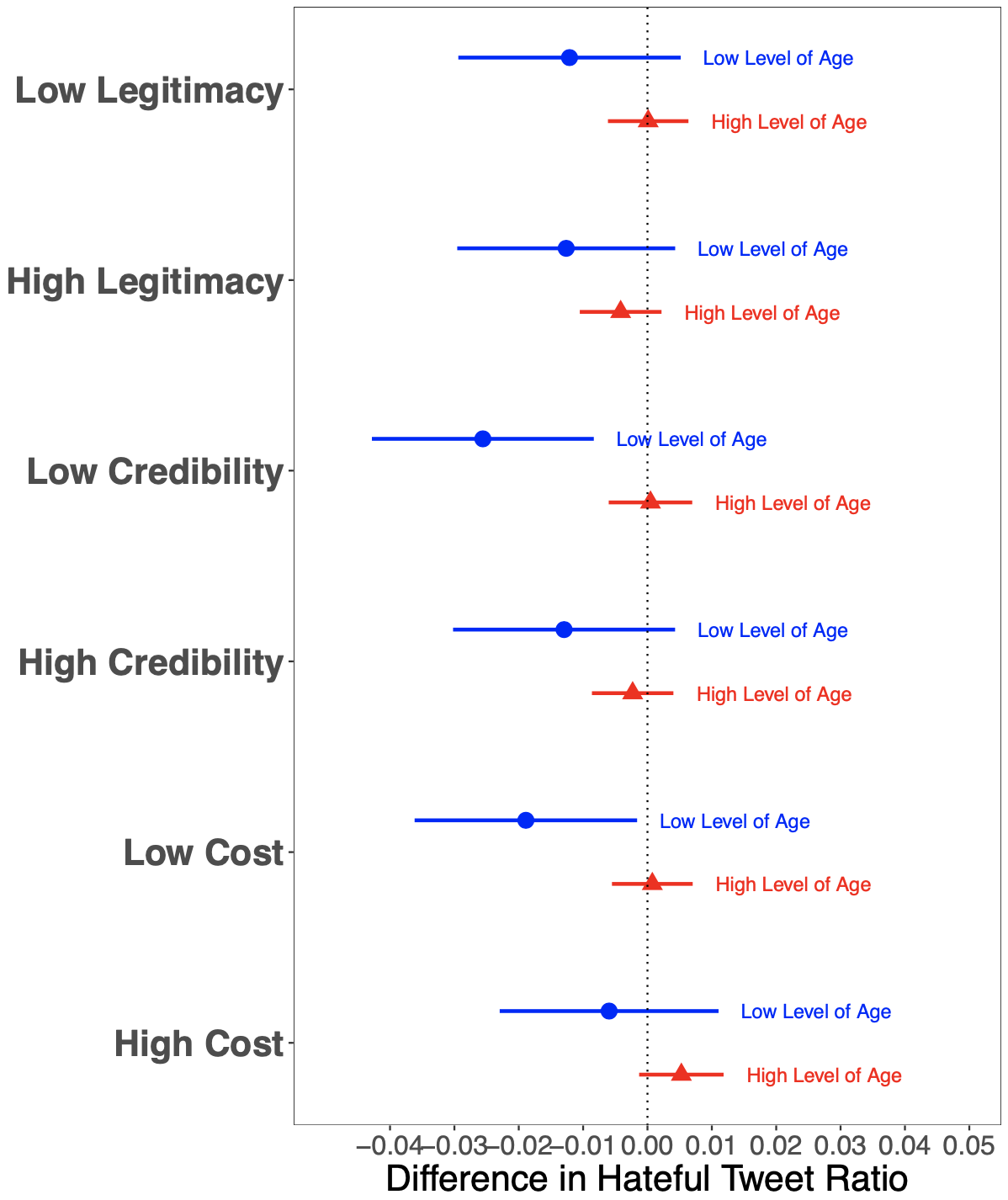
As the findings in Figure C2 show, nor do we find any support for HE 2, in the sense that the effects of warning tweets do not increase in magnitude when the users are less anonymous. Interestingly, however, the negative effects of the high legitimacy and high cost tweets that we observe in figures 3 and 4 seem to be driven by the users who have either their photos or their name in their profile but not both of them.

Below, we show whether having an old or young profile, or whether being more or less active on Twitter affects the effectiveness of our treatments.



**Figure C3**

The coefficient plot in Figure C3 shows that as opposed to what we expect, the decrease in the use of hateful language seems to be driven by users who are comparatively less active in Twitter, as measured by the number of statuses they post over the month prior to our treatment. This is especially true for the effects of the high legitimacy tweet. This could be because more active users might feel confident that given their heavy usage, and given that they have not got into trouble in the past, they are not likely to get suspended because of our warnings. This explanation is directly tied to the importance of "personal experiences with punishment and punishment avoidance (Stafford and Warr, 1993, 127)", suggesting that general deterrence may be working next to specific deterrence.

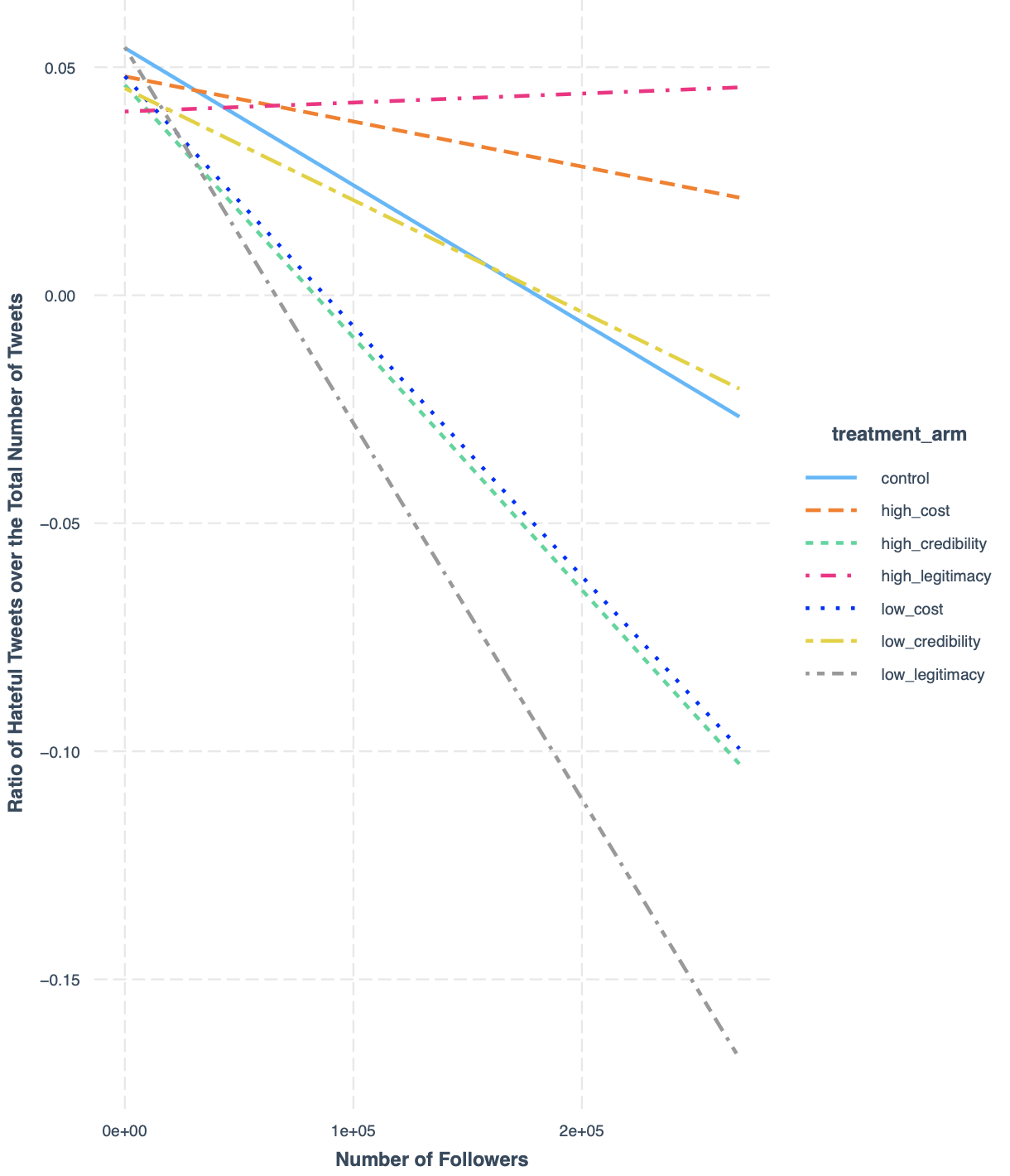


**Figure C4**

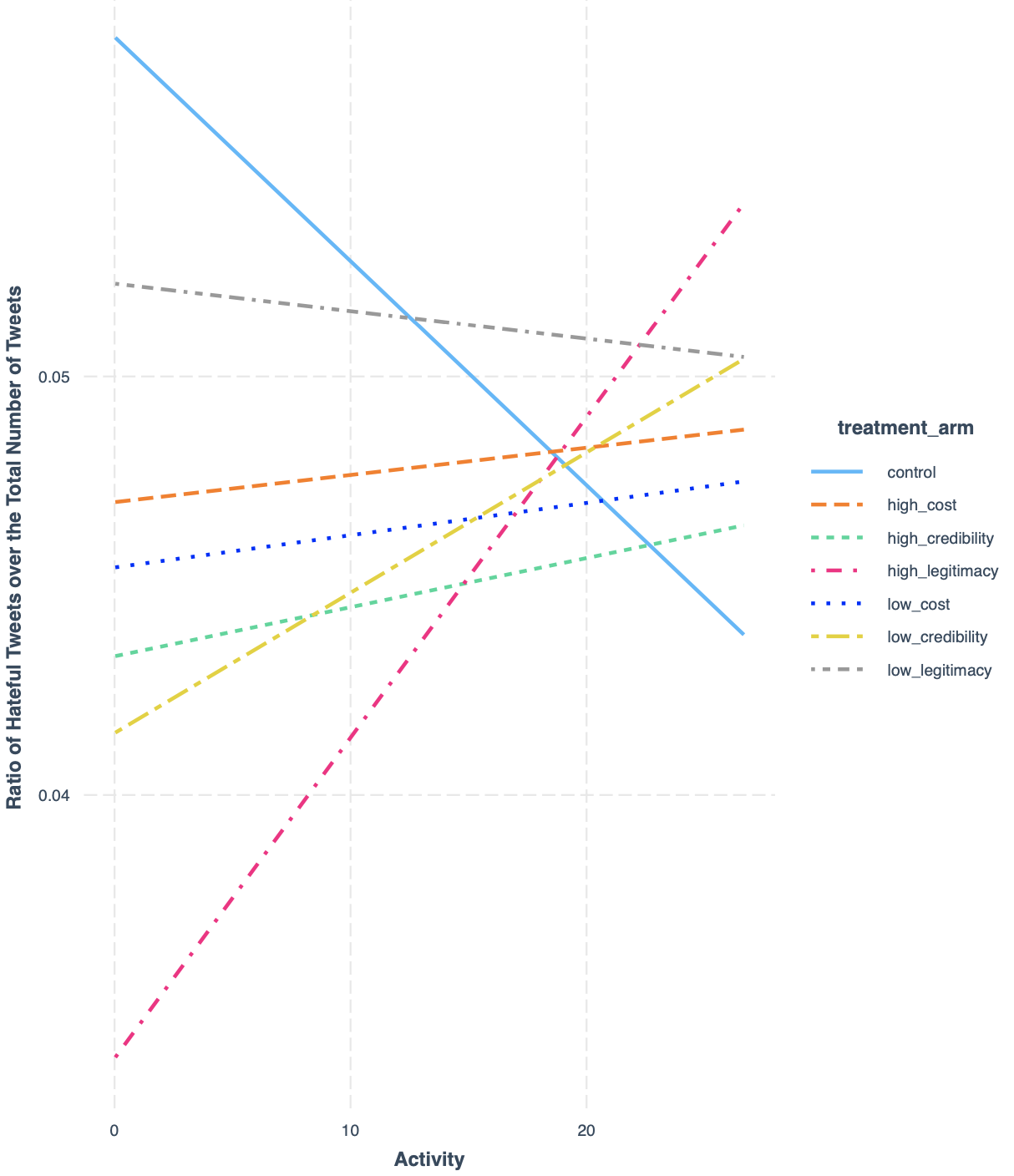
Similarly, the coefficient plot in Figure A7 shows that as opposed to what we expect, the decrease in the use of hateful language seems to be driven by users who have comparatively newer Twitter profiles, as measured by the log number of days since the creation of their profile. This is especially true for the effects of the high legitimacy tweet. Both Figure A6 and A7 show if a user is more entrenched in Twitter by being more active and for a longer time period on the platform, rather than becoming more careful with their language after receiving a warning, they tend to be desensitized to such warnings.

**Appendix D: Heterogeneous Treatment Effects with Continuous Variables**

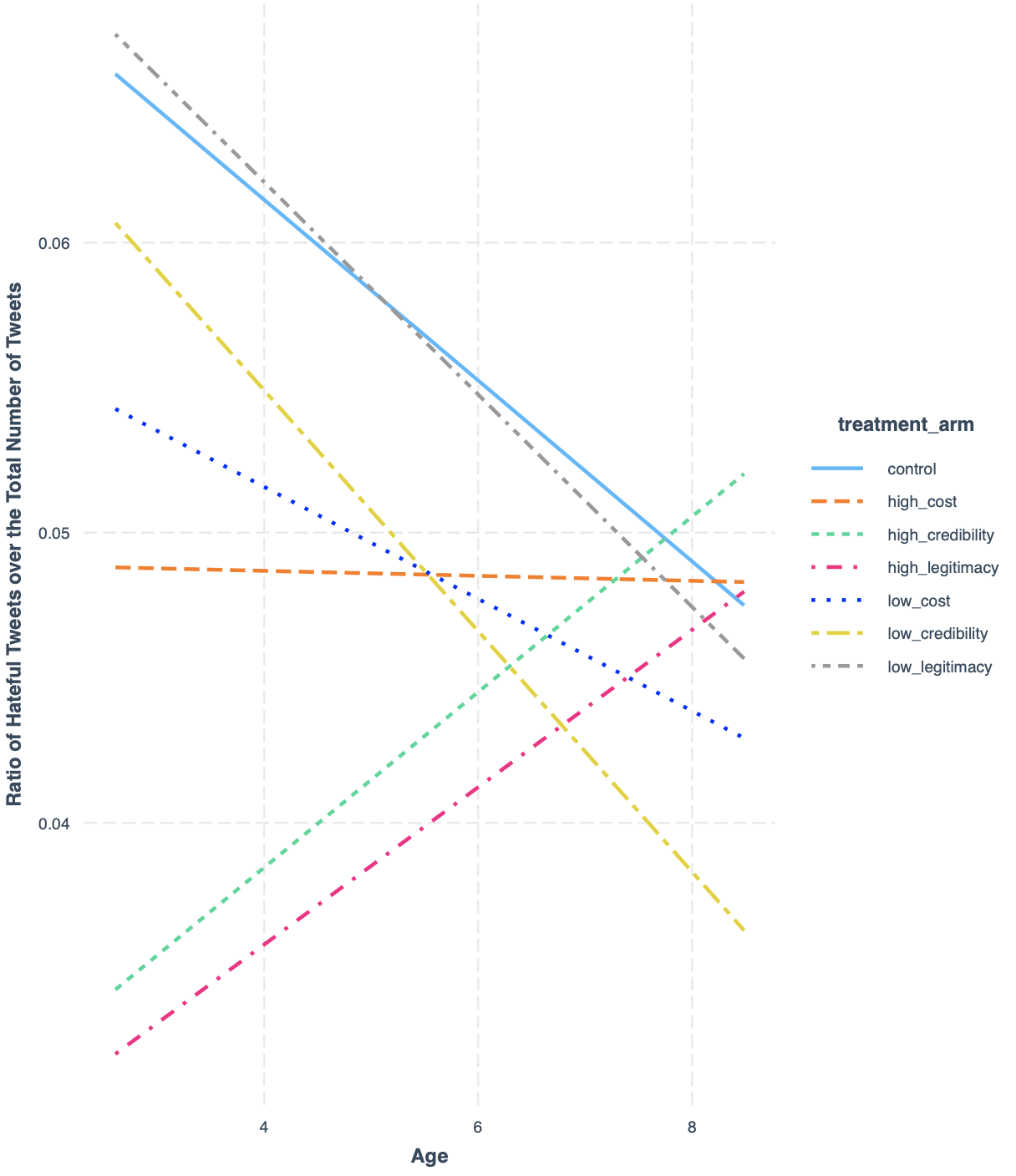
In our PAP, we hypothesized that the interaction effects with the variables number of followers, pre-treatment activity level, profile age, and profile anonymity would amplify the effect of our treatments. However, none of the findings from the interactions provide support for our hypotheses, as they are all insignificant. Below, Figures A7-A10 show prediction plots for the interaction effects.



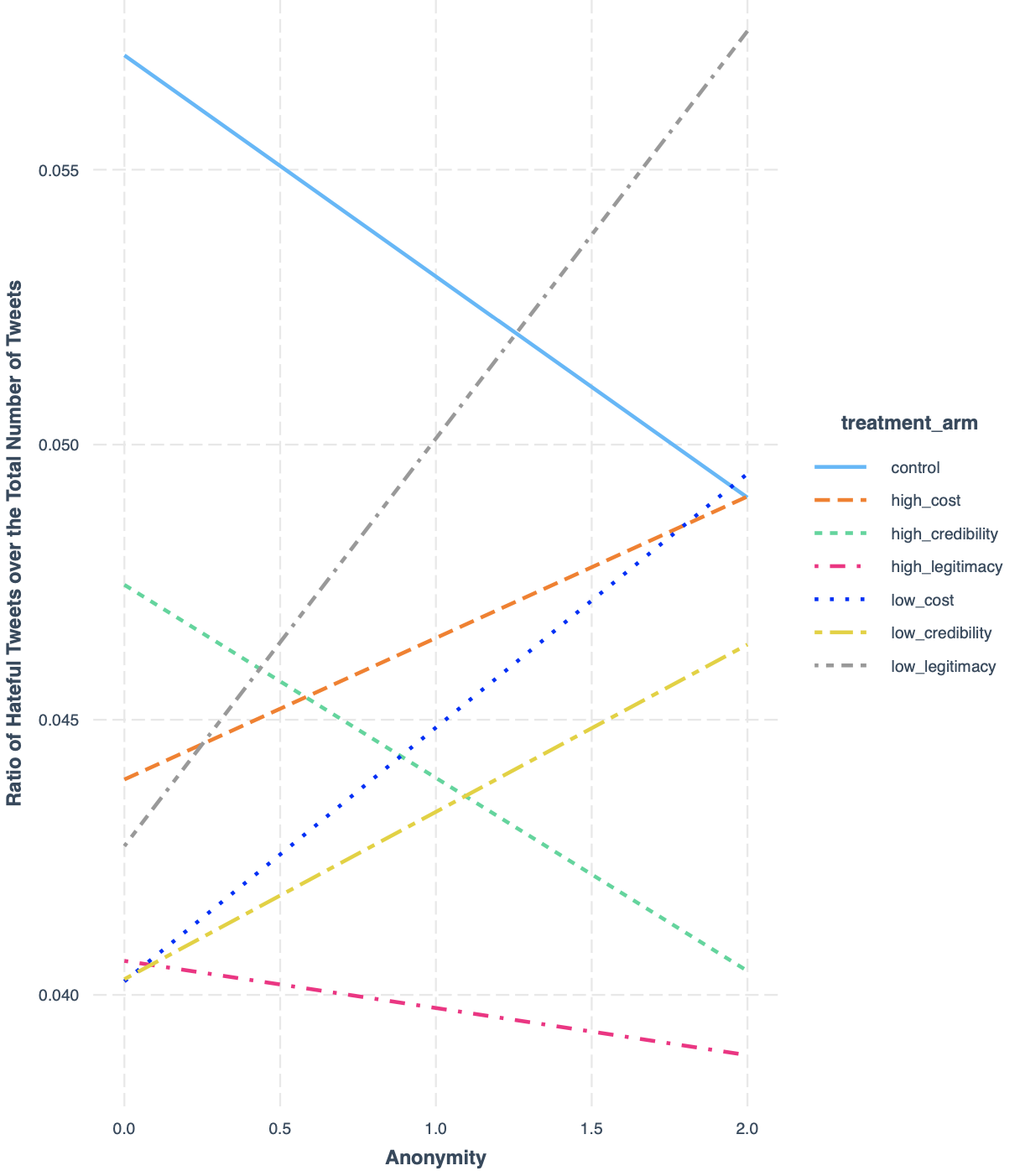
**Figure D1**

****

**Figure D2**

****

**Figure D3**

****

**Figure D4**

**Appendix E: Manipulation checks for our warning tweets**

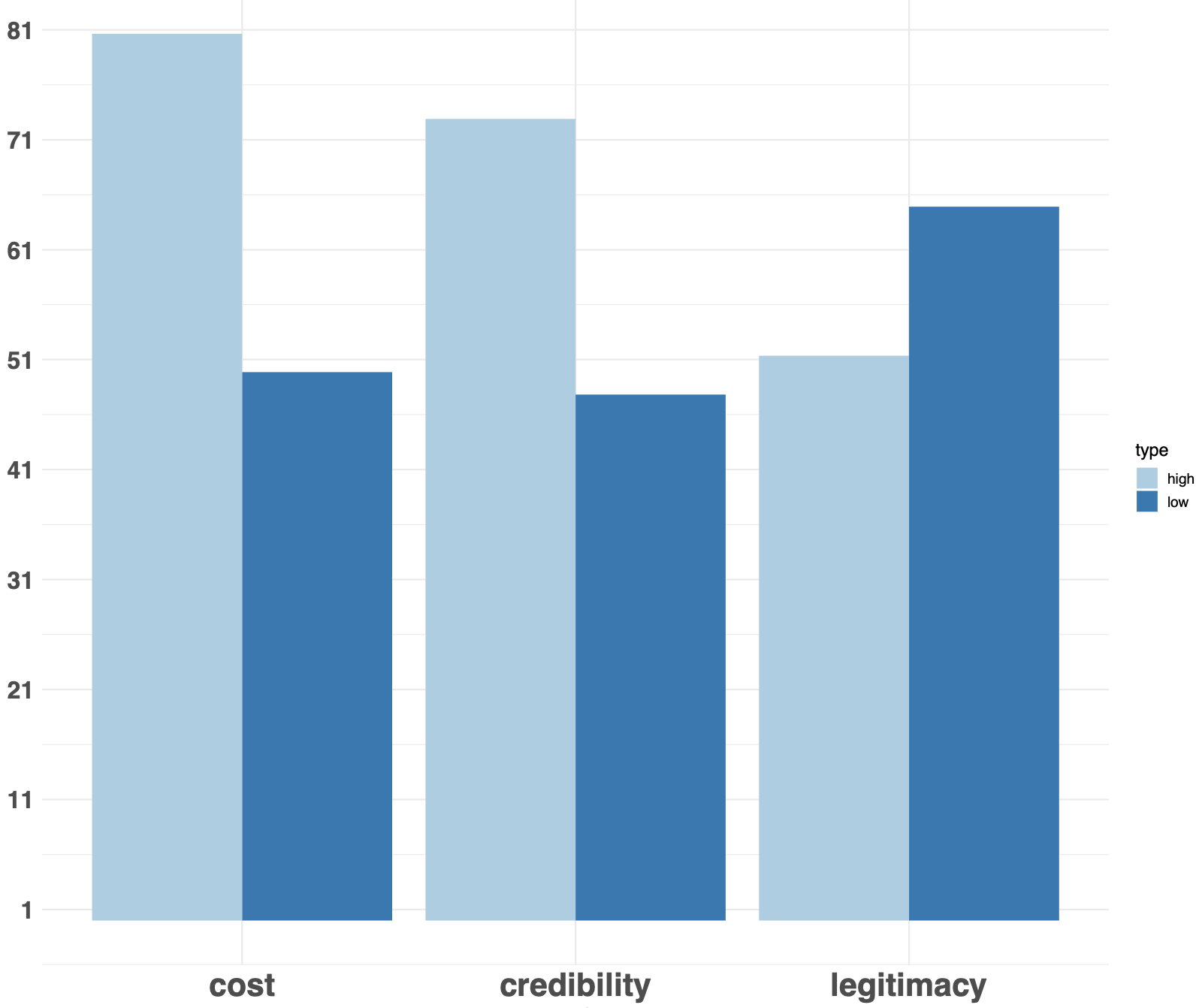
We design our tweets along three dimensions: costliness, credibility, and legitimacy. For each dimension, we designed survey questions that test whether we are able to manipulate what we aim to manipulate with each warning tweet.

For costliness tweets, we presented the respondents with the high and low cost tweets that we prepared. We then asked them the following question: “If you had used hateful language on Twitter and received this tweet how costly do you think it would be if you lost your Twitter account?” We then asked the respondents to give a score from 0 to 100 to each tweet in terms of how costly they think they would consider their suspension. The higher a given tweet is scored, the higher the costliness of the tweet is.

For legitimacy tweets, we presented the respondents with the high and low legitimacy tweets that we prepared. We then asked them the following question: “If you had used hateful language on Twitter, would you have a problem with this user sending you this tweet?” We then asked the respondents to give a score from 0 to 100 to each tweet in terms of how much problem they would have receiving it. The higher a given tweet is scored, the lower the legitimacy of the tweet is.

For credibility tweets, we presented the respondents with the high and low credibility tweets that we prepared. We then asked them the following question: “If you had used hateful language on Twitter, would you believe that you might get suspended after receiving this tweet?” We then asked the respondents to give a score from 0 to 100 to each tweet in terms of how convincing they would find their potential suspension due to their use of hateful language. The higher a given tweet is scored, the higher the credibility of the tweet is.

We ran this survey with 50 respondents from Mturk. We present the results below:



**Figure E1**

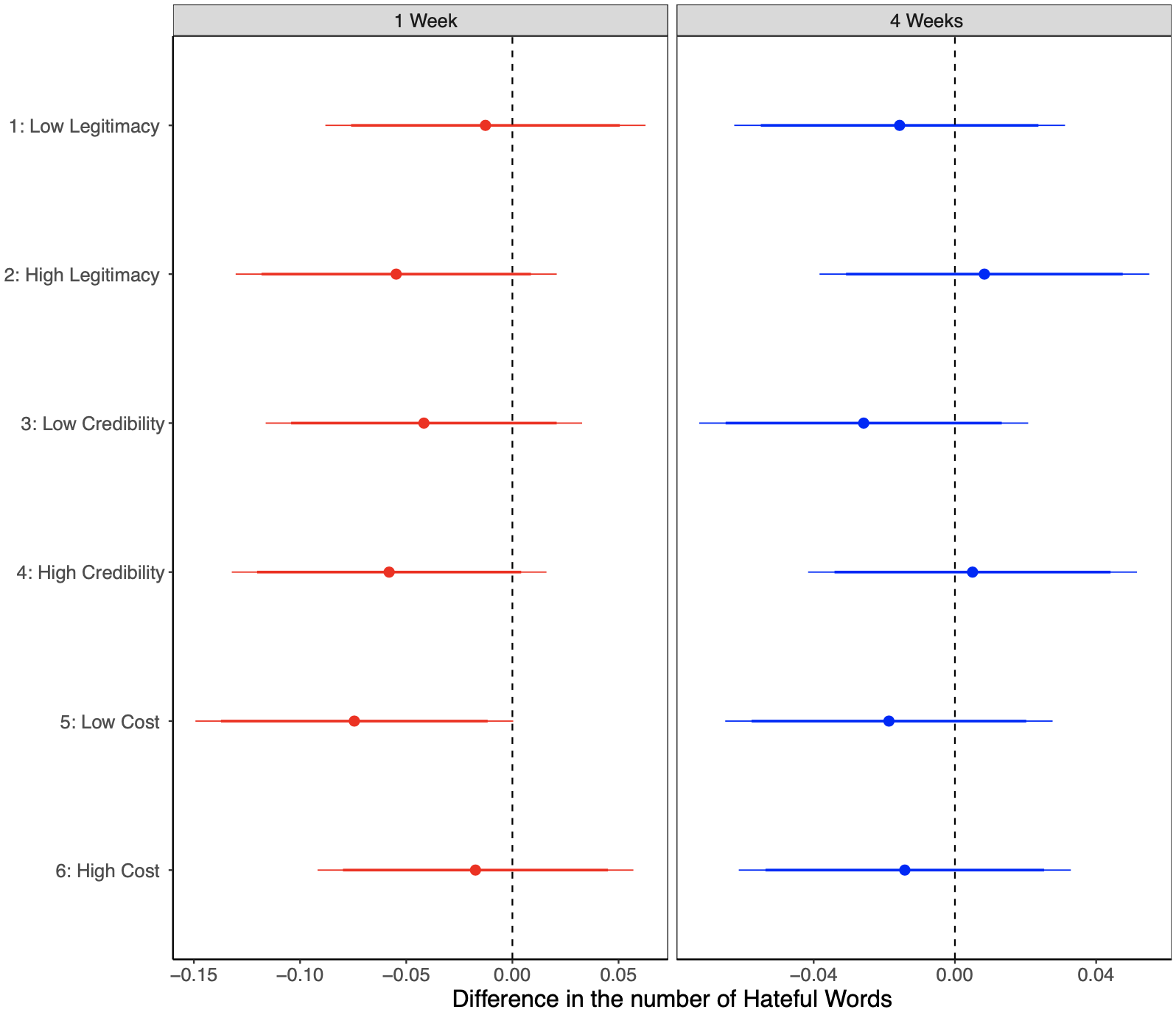
On average, respondents gave a score of 81 to high costliness tweets, whereas they gave only 50 to low costliness tweets. They gave a score of 72 to high credibility tweets, whereas they gave only 50 to low credibility tweets. They gave a score of 62 to low legitimacy tweet, whereas they gave only 51 to high legitimacy tweets (keep in mind that because of the way in which we asked our question about the legitimacy tweets, the direction of difference between the mean scores here is reversed).

**Appendix F: Alternative Outcome Variable**

In PAP, we state that we are going to measure our outcome variable, that is the use of hateful language, by taking the ratio of tweets that contain at least one hateful word over the total number of tweets a user tweets in a given period of time. The reason is that we want to see whether there is a decrease in the intention of tweeting a tweet that contains at least one hateful word in the first place.

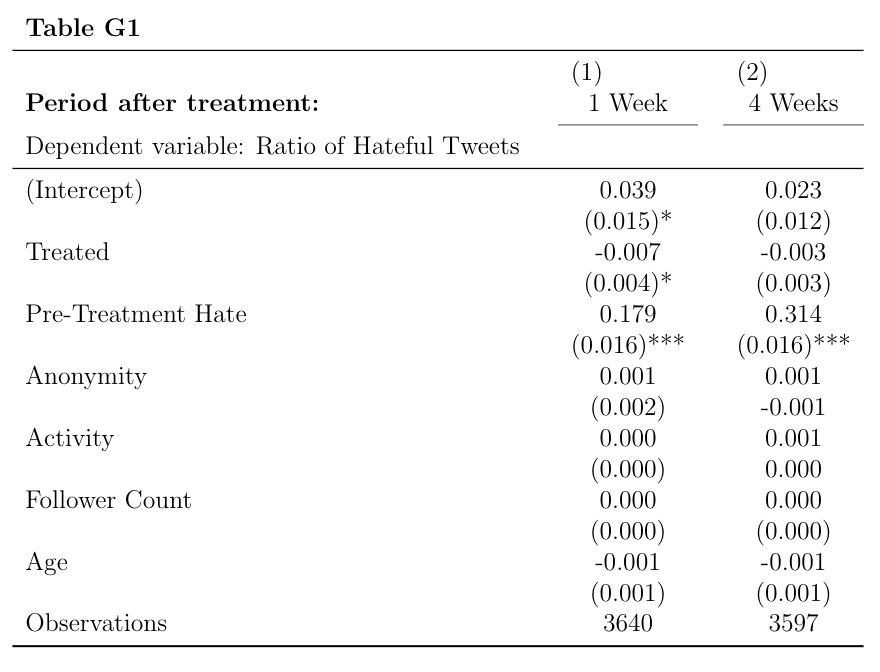
For transparency purposes, we report here that the pre-analysis plan had one inconsistency where one of the econometric specifications that we report does not match the description of the outcome variable that we detail much more in PAP in the variables section. As such, in the main paper, we focus on the outcome that was explained in the variables section of our PAP. We also deem that the outcome we report in the main paper is a better measure as it takes into account the overall volume of the tweets.

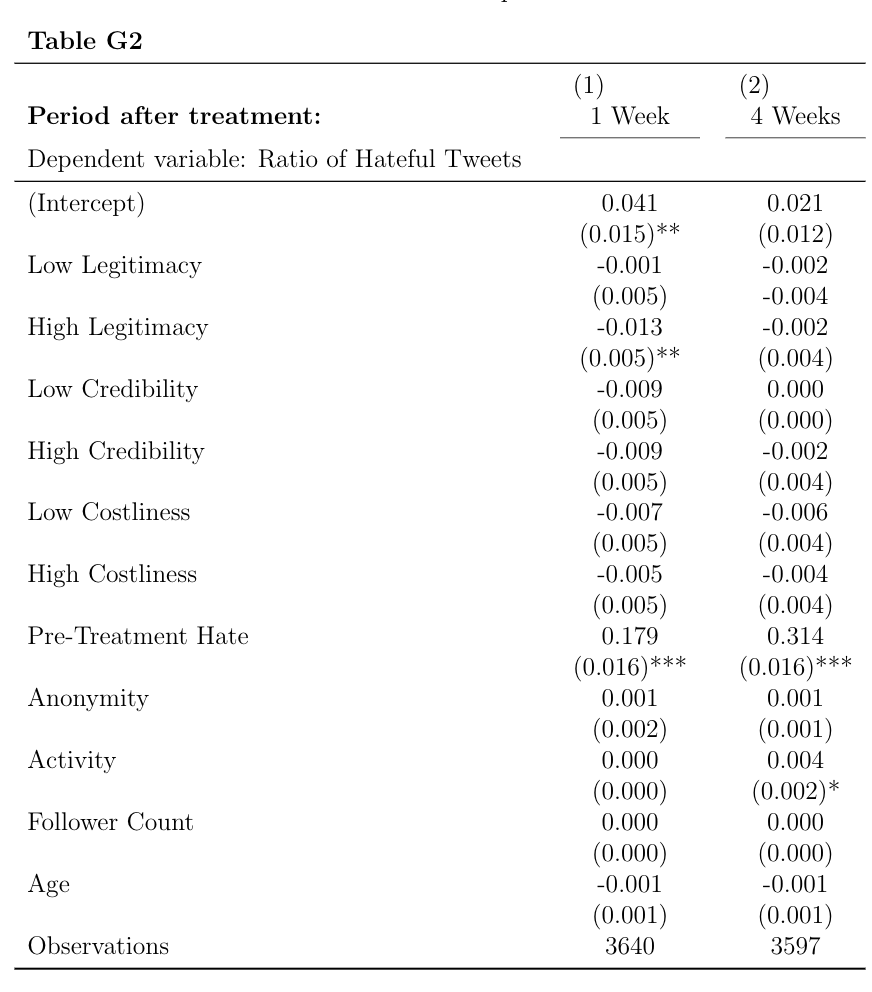
Figure F1 shows the alternative outcome that we report in PAP, that is the daily average number of hateful words. Instead of the high legitimacy tweet reducing the hateful language significantly, we see that the low cost tweet leads to a significant decrease in the use of hateful language.

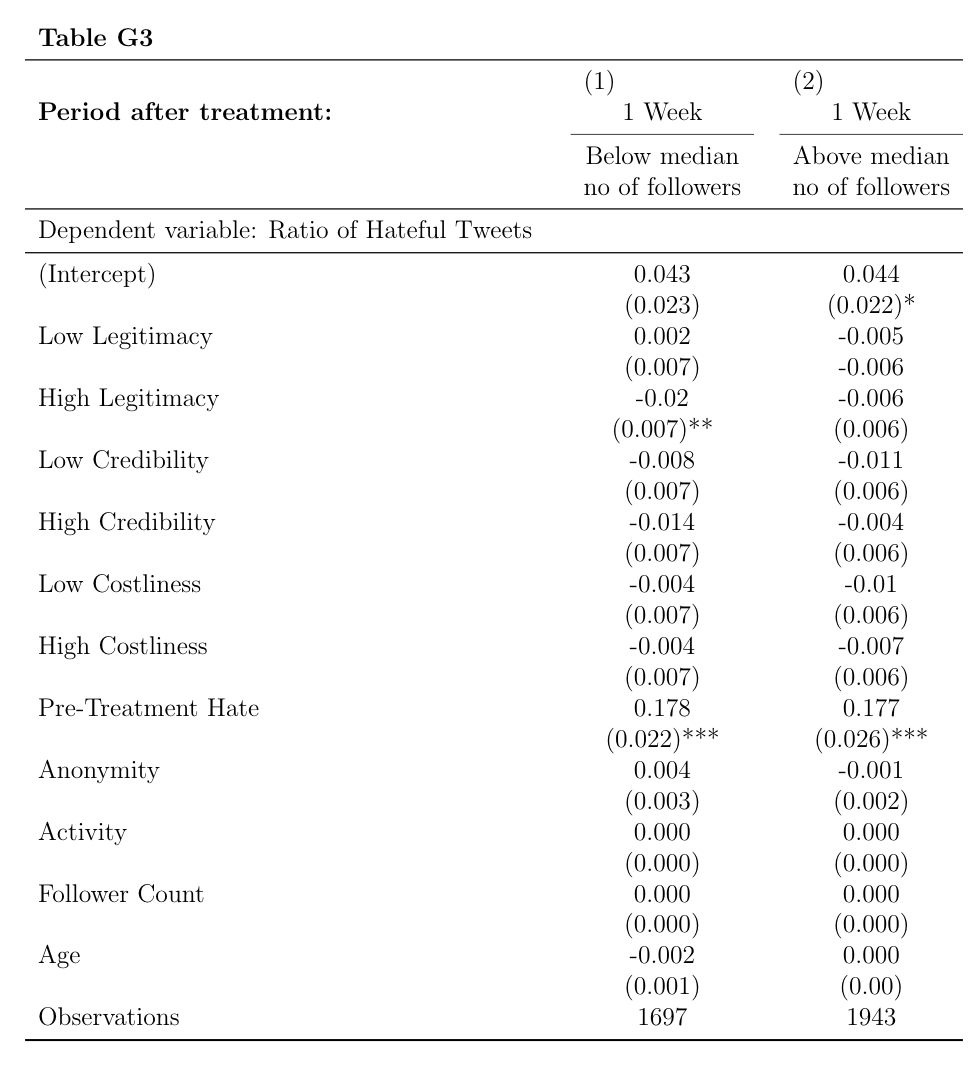


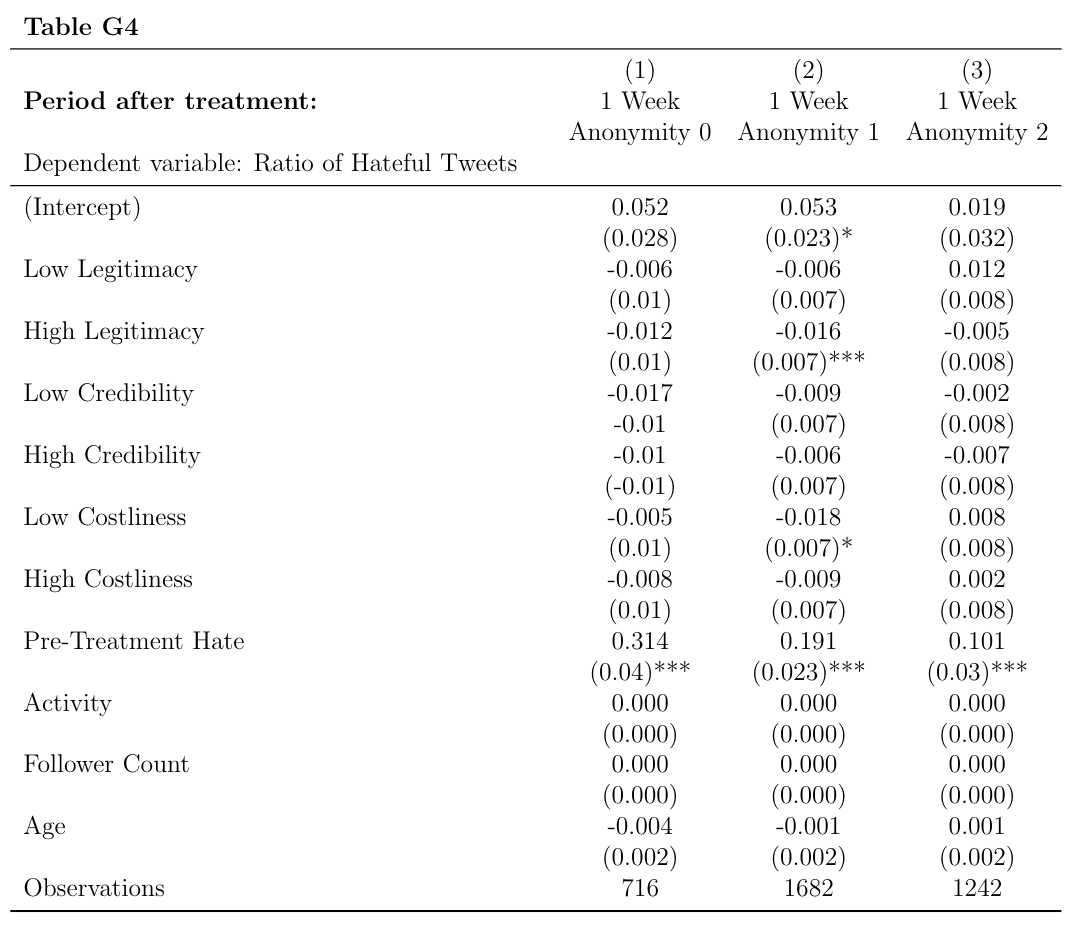
**Figure F1**

**Appendix G:** **Detailed Tables for the Figures in the Paper**

****







**Appendix H: Treatment Tweets and Profiles that Sent Them**

**Low legitimacy:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**High legitimacy:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**Low costliness:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**High costliness:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**Low credibility:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**High credibility:**

* **Tweet:**
  + ****
* **Profile:**
  + ****

**Appendix I: Pre-Registered Hypotheses and their Theoretical Justifications**

As we specify in Footnote 4, in the interest of space and at the advice of the editors, we do not provide the theoretical justification for testing the channels of costliness, credibility and legitimacy. In this section, we discuss each channel, and provide our hypotheses that we pre-registered prior to our experiment.

In addition to arguing that a warning message can lead to a decrease in the use of hateful language, we explore what makes warning messages deterrent. Scholars of deterrence make a conceptual distinction between deterrence and deterrability, where deterrability refers to the willingness of the offender to engage in weighing risks against rewards of their behavior (Pogarsky 2002, 431). Deterrability can change depending on the degree of objective deterrence, the actual level of punishment, or depending on the degree of perceptual deterrence, which emphasizes the individual’s perception of the threat of punishment (Waldo and Chiricos, 1972). Here, we explore three channels through which we can manipulate perceptual deterrence.

The first channel is costliness. Classic theories on deterrence emphasize the influence of the perceptions of sanctions’ severity in generating effective deterrence (Cusson 1993; Gibbs 1975; Paternoster 1987). Perceived severity means user’s conviction regarding the seriousness of the consequences associated with their behavior (Rogers 1975, 102). Therefore, if we can increase individuals’ understanding regarding the consequences of their deviant behavior, we can motivate their compliance with our warning messages. As Rogers (1975, 94-95) states, compared to low fear appeals, high fear appeals facilitate attitude change, which will also facilitate behavior change (Conzola and Wogalter 2001, 312). What differentiates high fear appeals from low fear appeals is the extent to which they emphasize an individual’s perception on how costly the consequences of their behavior are. This is important because individuals will not commit a crime if its costs are higher than their benefits (Beccaria 1963 [1764]).

If we apply this to the context of social media, or more particularly to Twitter, we can argue that warning tweets that emphasize the severity of account suspension should deter users from using hateful language more compared to tweets that do not emphasize this as much. One way to achieve costliness is to enhance users’ understanding as to what they will lose if their account gets suspended. Hence, we make the following hypothesis:

* H2: The more the warning tweet pushes the user to perceive their potential suspension as costly, the more likely it is that the user will decrease their use of hateful language.

The second channel is credibility. For a sanction to be effective, it needs to be credible (Nagin 1998, 8). Similarly, some scholars on deterrence argue that perceived certainty of sanctions can be more effective in deterring crime than the severity of sanctions (Pogarsky 2002, 432). In other words, the user’s conviction regarding the probability that a threat will occur affects how much they will be deterred from their unwanted behavior (Rogers 1975, 97). Hence, the user needs to be convinced and aware that they will face consequences as a result of their behavior. One factor that can make warnings credible is the authority of the source. Kiesler et al. (2012, 133-134) point out that moderation attempts on online platforms by members who seem to deserve to be in the moderator position are considered as being more credible by other members.

Based on these works from the literature on deterrence, we similarly argue that our warning tweets will deter users more from using hateful language if they are convinced that they might get suspended because of using hate speech. As the certainty of their potential suspension increases, they perceive their behavior as riskier (Pogarsky 2002, 432), which leads them to be more careful in their language to avoid the costs associated with the account suspension. Hence, we propose the following hypothesis:

* H3. The more credible the warning tweet is in terms of convincing the user that they might get suspended as a result of using hateful language, the more likely it is that the user will decrease their use of hateful language.

The third channel is the legitimacy of the warning message. Sherman (1993, 445) argues that the legitimacy of experienced punishment is essential for the acknowledgement of shame, which then conditions deterrence. Along the same lines, Tyler (1990) distinguishes between sanctions citizens perceive as fair or unfair. Fair sanctions lead to deterrence whereas unfair sanctions do not. When it comes to the connection between the legitimacy of a warning message and its deterrent effects, we employ the mechanism that is put forward by Scheff and Retzinger (1991), which is later embraced by Sherman (1993, 459) when he explains how a legitimate punishment is more likely to deter the deviant behavior. When a warning message lacks legitimacy, the target feels disrespected, which makes them ashamed of being ashamed, “infusing them with self-righteousness” (Sherman 1993, 447). This then makes the warning ineffective in deterring unwanted behavior. The mechanism is also in line with Durkheim’s hypothesis that a punishment has an influence only if the person whom it is inflicted on does not feel disrespected (Durkheim 1961). Therefore, we argue that the legitimacy that a target grants to a warning message is driven by the warner’s respectfulness (Sherman 1993. 459).

Hence, on Twitter, a warning tweet would be more effective the more legitimate it is. In other words, it would deter the users from using hateful language more the more it respects the target. The more the target perceives receiving a tweet that warns them of suspension because of their behavior as acceptable, the less likely they are to feel disrespected, and the more likely they are to reduce their use of hateful language. Therefore, we make the following hypothesis:

* H4. The more the user considers being sent a warning tweet as something acceptable (i.e. as legitimate), the more likely it is that the user will decrease their use of hateful language.

**Heterogeneous treatment effects:**

In addition to different mechanisms through which a warning message of suspension can reduce the use of hateful language, we hypothesized that the impact of the warning might depend on characteristics of the target population, or, put another way, that there might be heterogeneous treatment effects. If the target population is such that they will lose more as a result of suspension, then the warning message should have a higher deterrent effect for them. On Twitter, it takes time for a user to obtain a list of friends that is in line with their tastes, to establish their follower base and to post tweets that sustain this base. The number of followers a user has, for example, is so important on Twitter that there are studies who predict NBA players’ salaries based on their follower base on Twitter (Li and Huang 2014). However, once a user gets suspended, they lose access to all of their followers, friends, and previous posts. In order to get their followers back, they need to build a new profile from scratch, and reach out to their followers to inform them that their previous account got suspended. Hence, the more they invest in their profile as measured by the tweets they post, the number of followers and friends they have, the more they have to lose from a suspension of their account. Therefore, we hypothesize the following:

* HE 1: Our warning messages will reduce the use of hateful language more for people who have

a) more followers,

b) older profiles,

c) who are more active on Twitter.

Another characteristic of the target population that would make our warning tweets more or less effective is their anonymity level. Cybercrime scholars argue that due to the anonymous nature of cyberspace, the perceived certainty of online punishment is low (Andress and Winterfeld, 2011, 250). Consequently, offenders’ perceived risk of detection is low (Geerken and Gove 1974 498), which then decreases the threat of sanctions (Denning and Baugh, 1999; Harknett, 1996; Harknett, Callaghan, and Kauffman, 2010).

Applying the same logic to Twitter, we can argue that users who are anonymous (i.e. users who do not reveal their names or photos in their profile) would be less sensitive to our warning messages because anonymous users’ perceived risk of detection would be lower (Munger 2017, 630-631). Moreover, users who put their names or photos in their profiles would be more likely to think that the online consequences that we warn them of could also affect their offline life. Evidence corroborating this can be found in a study by Omernick and Sood (2013), who show that anonymous users provide lower quality comments on online platforms compared to non-anonymous users. Hence, we make the following hypothesis:

* HE 2: Our warning messages will reduce the use of hateful language more for people who have less anonymous profiles.

Appendix J: Dictionary of Hateful Words

“beaner”,“chinc”, “chink”,“coon”, “dego”, “gook”, “guido”, “heeb”, “kike”, “kyke”,

“jigaboo”, “mick”, “negro”,“nigger\*”, “niglet”, “paki”, “porchmonkey”, “pollock”,

“ruski”, “sandn\*\*\*\*r”, “spic”, ”wop”, ”jap”, “junglebunny”, ”spick”, “wetback” “bitch”,

“cunt”, “dyke”, “skank\*”, “slut\*”, “whore\*”, “ho” “\*bastard\*”, “\*shit\*”, “\*fuck\*”,

”carpetmuncher”, “\*cock\*”, “cum”, “\*douche\*”, “\*fag\*”, ”fudgepacker”, “blowjob”,

“handjob”, “homo”, “jizz”, ”lesbo”, “lezzie”, “pussy”, ”queerbait”, ”rimjob”, “skeet”,

“tard”

1. In the PAP, we indicated that we were going to treat our moderator variables as continuous variables. However, upon rethinking the visualization qualities, we have chosen instead to present the results here using discrete variables because they are easier to interpret; please see the appendix for the presentation of results using continuous versions of these variables as specified in the PAP. Also, please see the appendix for other variables that we use to test HE1 such as the age of a user’s profile and their pre-treatment activity level. [↑](#footnote-ref-0)
2. Please see Table G3 in Appendix G for more details on sample size and control coefficients. [↑](#footnote-ref-1)
3. Please see Table G4 in Appendix G for more details on sample size and control coefficients. [↑](#footnote-ref-2)