Gov 2001: Replication Report

Mohammed Alsobay, Sumaya Malas, Tianyu Qiao (Team 2.5 Arabs)

11/8/2021

Overview

The paper we are replicating is "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online." The main point of the paper is that different types of "priming", oriented around religious or political identity, can decrease sectarian hate speech and tolerance thereof. This is demonstrated in two ways: a Twitter "experiment" and a survey experiment targeting two different audiences, in which elite-supported messages of common identity affected the outcome of hostile outgroup sentiment. For the first experiment, the authors would reply to Arab Twitter users who had regularly tweeted hostile sectarian language with a randomly assigned counter-speech message and measured the post-treatment over time. The second experiment was collecting survey data in Lebanon and suggested that if respondents were primed with messages of common-religious-identity they rated sectarian tweets more negatively and counter-sectarian tweets more positively.

In the paper's main text, there are 7 figures and no tables. Here, we replicate the figures (excluding Figure 1, which is a screenshot of Twitter), and verify some of the measures and claims made in text. The replication is executed by running the code provided by the authors "as-is" and, with a few exceptions, the figures *do* replicate. For each figure, we present the replication followed by the original version in the paper. We discuss the exceptions and how they might have arisen in more detail within the relevant sections. Finally, we discuss possible directions for our extension of this paper.

Comments on in-text data descriptions

- The paper mentions "we treated 50 subjects every day for 20 days between January 31, 2018, and February 19, 2018, for a total of 9,957 subjects." This is likely a slip of the finger, as the calculation described gives 1000 subjects, and the number of subjects in the data is 957.
- In describing a sub-analysis, the exclusion criterion is described as "we restricted our analysis to users who have the median number of followers (250) or fewer". In the code provided by the authors, the threshold is 245 rather than 250, and 245 is in fact the median measured on the data.

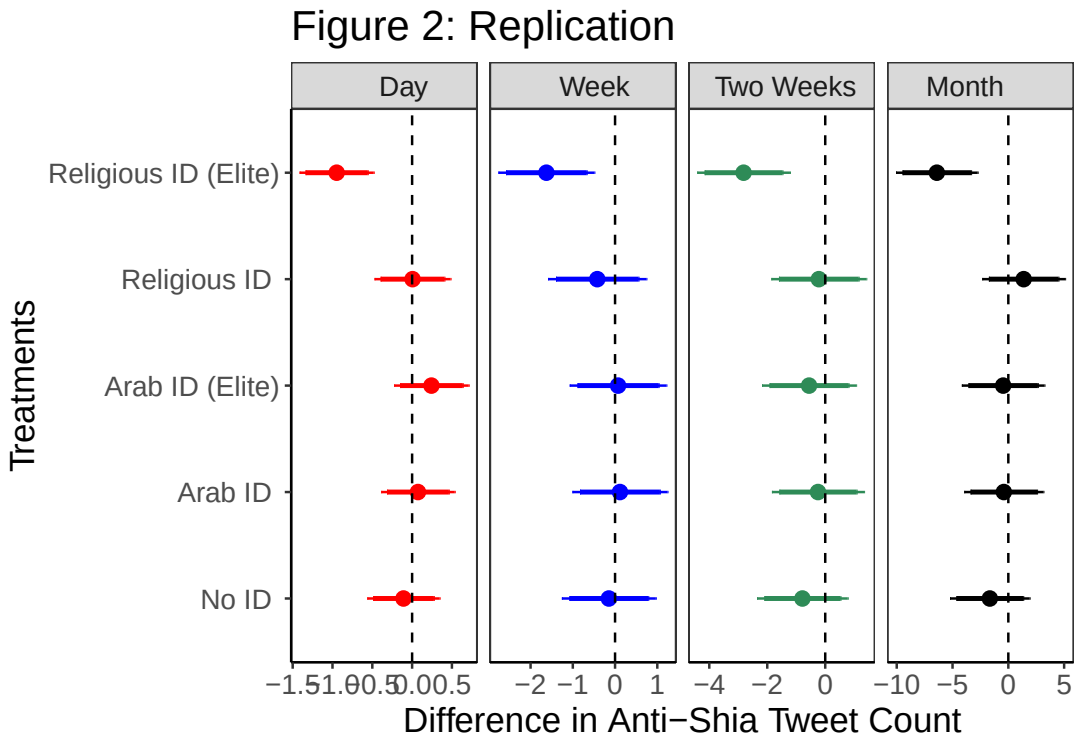
Figures Pt. 1: Twitter Experiment

Figure 1 is a screenshot of the sockpuppet Twitter account, so we begin from Figure 2. The pre-processing code has been excluded from this report for brevity, and is mostly creating subsets of the data for the various figures (e.g. high anti-shia friends, below median number of followers, etc.).

Figure 2: Effect of Treatment on Volume of Anti-Shia Tweets

This figure is replicated from the authors' code successfully.

```
#Difference in Means (All Data)
month<-lm(month_anti_shia_post-month_anti_shia_pre ~ treatment_num, data=data)
two_weeks<-lm(two_weeks_anti_shia_post-two_weeks_anti_shia_pre ~ treatment_num,
              data=data)
week<-lm(week_anti_shia_post-week_anti_shia_pre ~ treatment_num, data=data)
day<-lm(tpd_anti_shia_post-tpd_anti_shia_pre ~ treatment_num, data=data)


multiplot(day, week, two_weeks, month,
          coefficients=c("treatment_num1", "treatment_num2",
                         "treatment_num3", "treatment_num4", "treatment_num5"),
          newNames=c(treatment_num1="Arab ID ",
                     treatment_num2="Religious ID ",
                     treatment_num3="Arab ID (Elite)",
                     treatment_num4="Religious ID (Elite)",
                     treatment_num5=" No ID "),
          names=c("     Day", "   Week", "  Two Weeks ", "Month    "),
          title="Figure 2: Replication", scales="free_x",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen", "black")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(),
                     legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=15))+
  ylab("Treatments") +
  xlab("Difference in Anti-Shia Tweet Count")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```




```
knitr::include_graphics(path="replication_figures/figure_2_orig.png")
```

FIGURE 2. Effect of Treatment on Volume of Anti-Shia Tweets

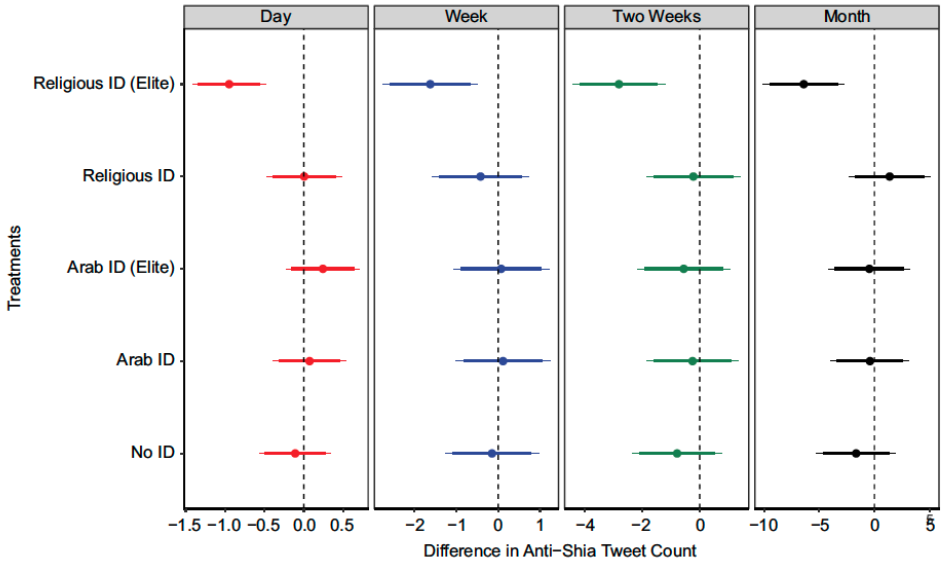


Figure 3: Distribution of Follower Counts

This figure fails to replicate entirely, and we end up with three versions of it:

1. The version in the paper, which conveys a slightly false distribution, described below, and is in greyscale.
2. Our independent replication, which agrees with the version in the authors' code (in magenta), and shows the correct distribution.
3. The version in the authors' aforementioned "Figures" folder, which conveys the correct distribution but has a different axis than (2).

We make the following observations about the version published in the text:

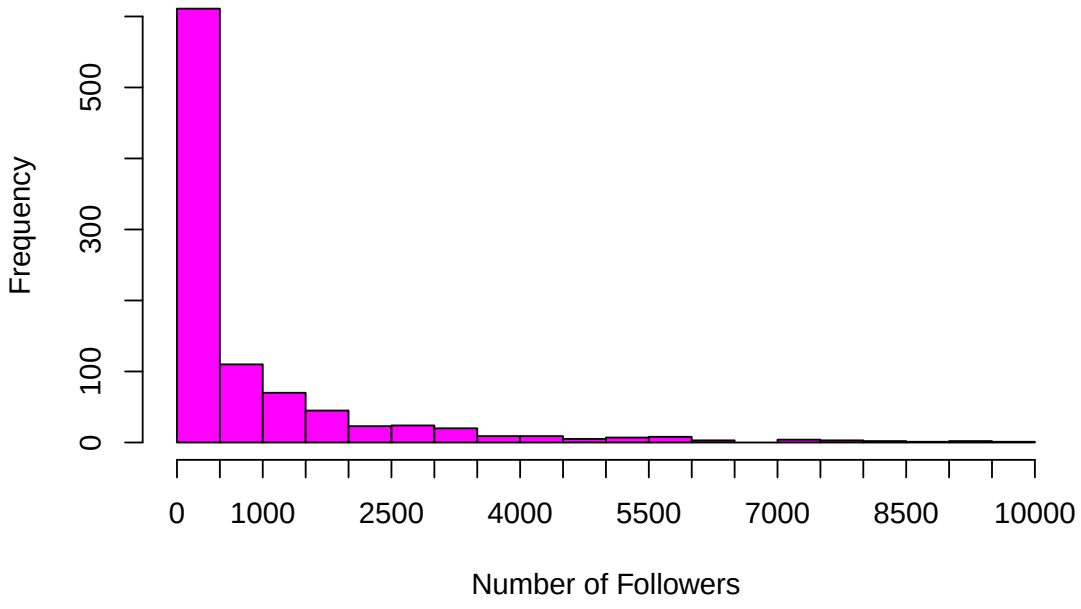
- The scale of each tick is inconsistent. 0 to 500 is 2 ticks, but 500 to 1500 is also 2 ticks, so is each tick 250 or 500 units? While bad practice in our opinion, it's entirely possible to generate this by manually specifying the "breaks" argument to the histogram function.
- A bit odder is that the tick labels don't align with the axis correctly, e.g. which ticks do 2500, 3500, and 4500 belong to?
- The counts don't match the actual raw data. In the data, there are 611 records with ≤ 500 followers, considerably less than implied by the figure (≈ 750).

Our guess for what happened here is:

1. They did something weird with the x-axis by accidentally applying numerical labels at points other than their actual value, which is why the counts in the published figure don't add up.
2. The code published to Dataverse is not the code for the final version of the paper, which is why even their code doesn't match the paper.

```
ticks<-seq(from=0, to=10000, by=500)
hist(data$followers_count, breaks=ticks, xaxt="n", xlab="Number of Followers",
     main="Figure 3: Replication", freq=TRUE, col="magenta")
axis(1, at=ticks)
```

Figure 3: Replication




```
knitr::include_graphics(path="replication_figures/figure_3_orig.png")
```

FIGURE 3. Distribution of Follower Counts

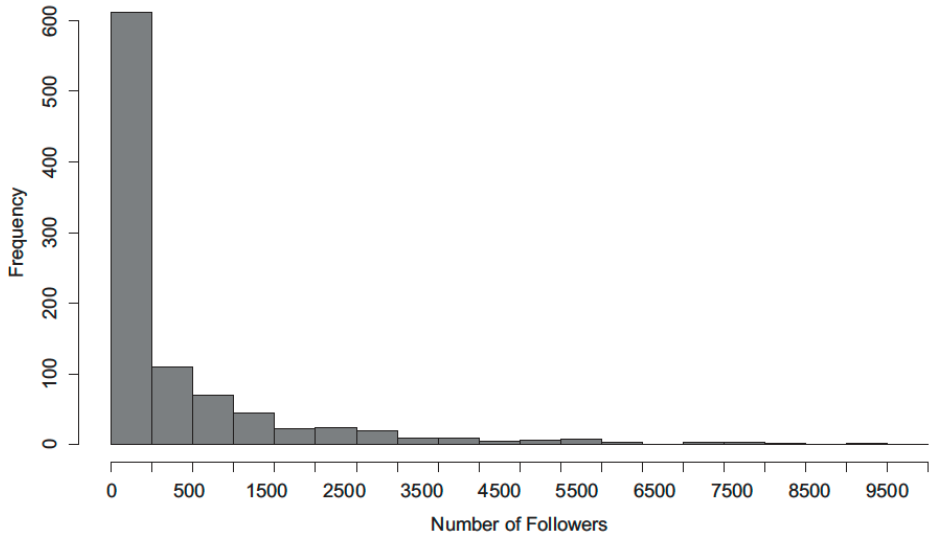


Figure 4: Effect of Treatment on Volume of Anti-Shia Tweets ($\leq$ Median Followers)

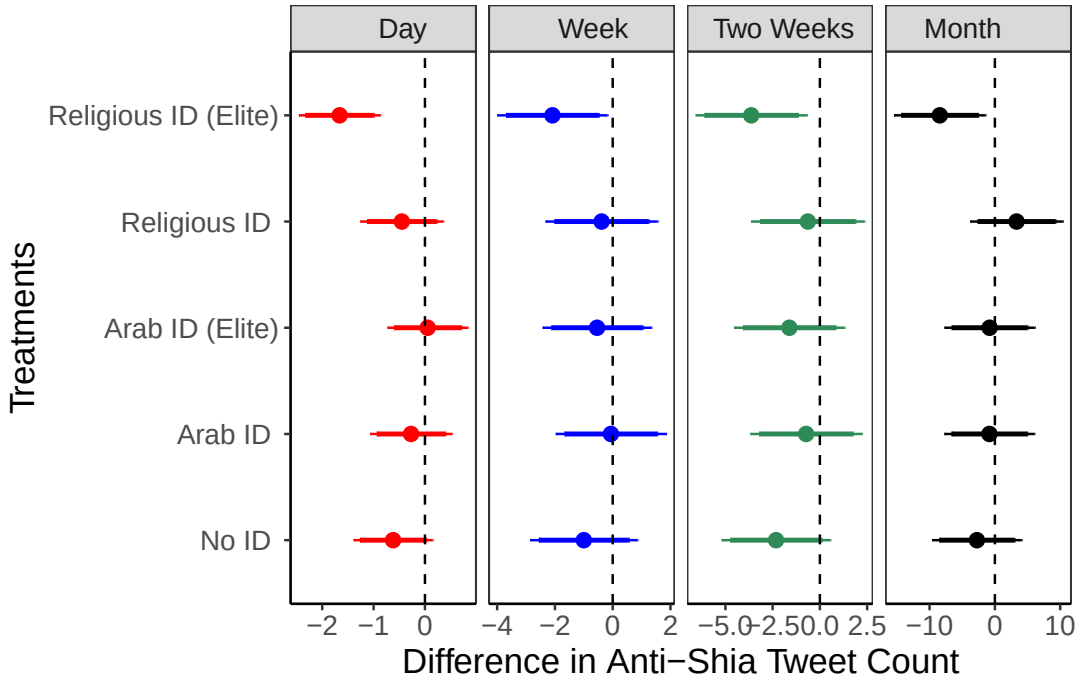
This figure is replicated from the authors' code successfully. To subset the data for this figure, they hard-coded the median number of Twitter followers as 245, which we confirm is the actual median in the data below. In our opinion, best practice would be to write it as a function of the data, so that it continues to be correct in the case of any data pre-processing.

```
median(data$followers_count)
```

```
## [1] 245
```

```
month<-lm(month_anti_shia_post-month_anti_shia_pre ~ treatment_num,
          data=data_median_fol)

two_weeks<-lm(two_weeks_anti_shia_post-two_weeks_anti_shia_pre ~ treatment_num,
              data=data_median_fol)

week<-lm(week_anti_shia_post-week_anti_shia_pre ~ treatment_num,
         data=data_median_fol)

day<-lm(tpd_anti_shia_post-tpd_anti_shia_pre ~ treatment_num,
        data=data_median_fol)

multiplot(day, week, two_weeks, month,
          coefficients=c("treatment_num1", "treatment_num2",
                         "treatment_num3", "treatment_num4", "treatment_num5"),
          newNames=c(treatment_num1="Arab ID ",
                     treatment_num2="Religious ID ",
                     treatment_num3="Arab ID (Elite)",
                     treatment_num4="Religious ID (Elite)",
                     treatment_num5=" No ID "),
          names=c("     Day", "   Week", "  Two Weeks ", "Month    "),
          title="Figure 4: Replication", scales="free_x",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen", "black")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(),
                     legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=15))+
  ylab("Treatments") +
  xlab("Difference in Anti-Shia Tweet Count")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```

Figure 4: Replication




```
knitr::include_graphics(path="replication_figures/figure_4_orig.png")
```

FIGURE 4. Effect of Treatment on Volume of Anti-Shia Tweets (≤ Median Followers)

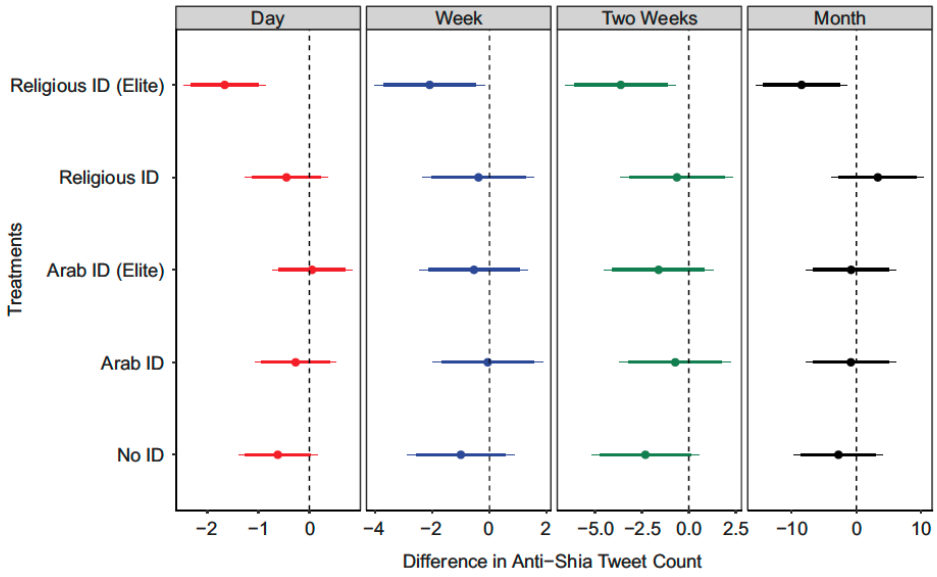


Figure 5: Effect of Treatment on Volume of Anti-Shia Tweets (Low vs. High Anti-Shia Friends)

This figure is replicated from the authors' code successfully. To subset the data for this figure, they hard-coded the median number of anti-Shia friends as 38, which we confirm is the actual median in the data below.

One slight discrepancy is that this figure is in color in both the replication code and the aforementioned "Figures" folder, and it's unclear why it ended up in greyscale in the actual paper – it is the only coefficient plot in the paper that deviates from the color scheme.

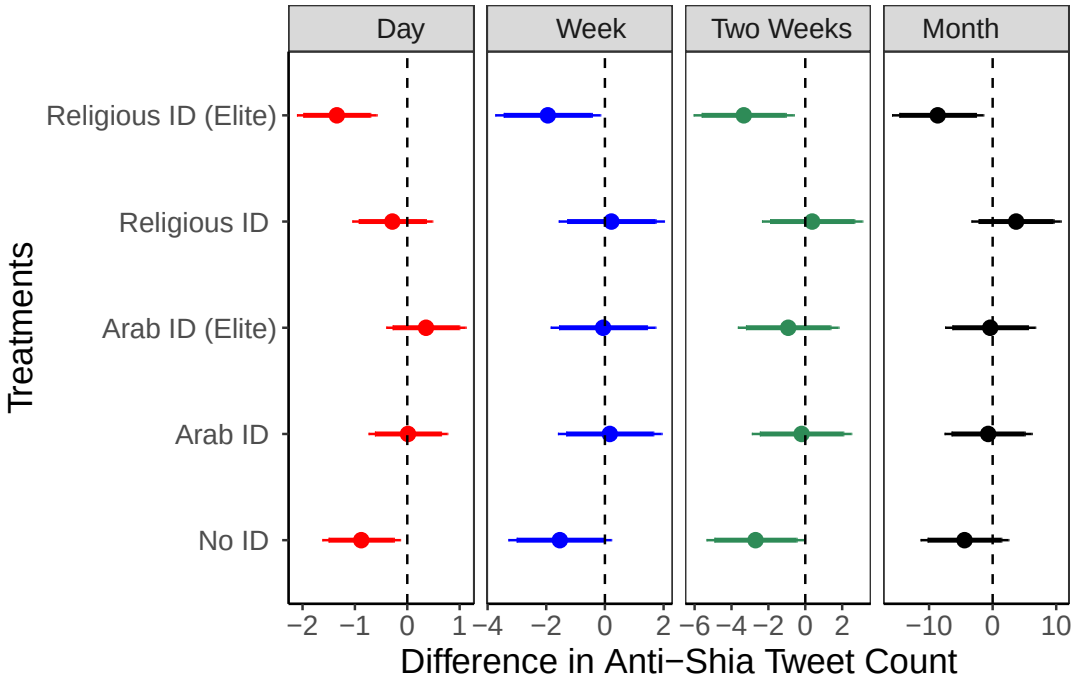
```
median(data$anti_shia_friends_count)
```

```
## [1] 38
```

5a: Low Number of Anti-Shia Friends

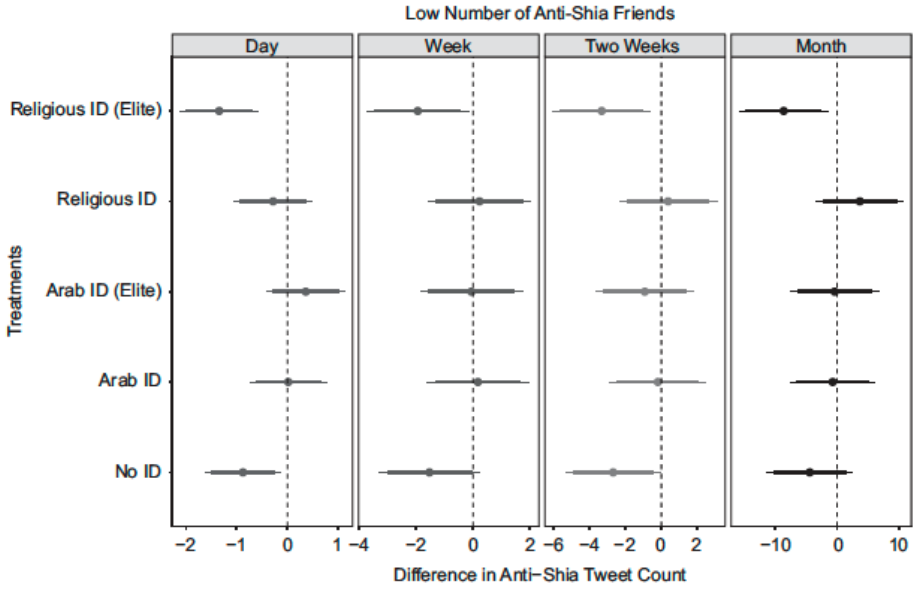
```
#Low Anti-Shia Friend Network
month<-lm(month_anti_shia_post-month_anti_shia_pre ~ treatment_num,
          data=data_anti_shia_net_low)

two_weeks<-lm(two_weeks_anti_shia_post-two_weeks_anti_shia_pre ~ treatment_num,
              data=data_anti_shia_net_low)

week<-lm(week_anti_shia_post-week_anti_shia_pre ~ treatment_num,
         data=data_anti_shia_net_low)

day<-lm(tpd_anti_shia_post-tpd_anti_shia_pre ~ treatment_num,
        data=data_anti_shia_net_low)

multiplot(day, week, two_weeks, month,
          coefficients=c("treatment_num1", "treatment_num2",
                         "treatment_num3", "treatment_num4", "treatment_num5"),
          newNames=c(treatment_num1="Arab ID ",
                     treatment_num2="Religious ID ",
                     treatment_num3="Arab ID (Elite)",
                     treatment_num4="Religious ID (Elite)",
                     treatment_num5=" No ID "),
          names=c("     Day", "   Week", "  Two Weeks ", "Month    "),
          title="Figure 5a: Replication", scales="free_x",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen", "black")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(),
                     legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=15))+
  ylab("Treatments") +
  xlab("Difference in Anti-Shia Tweet Count")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```

Figure 5a: Replication



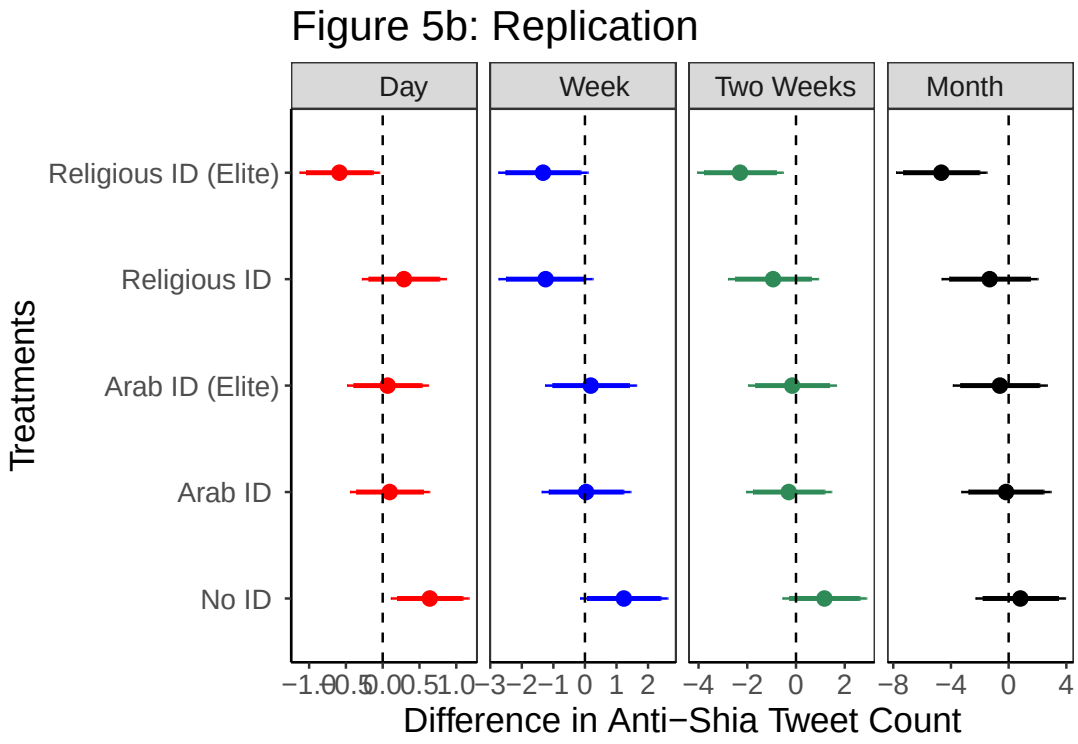

```
knitr::include_graphics(path="replication_figures/figure_5a_orig.png")
```

FIGURE 5. Effect of Treatment on Volume of Anti-Shia Tweets (Low vs. High Anti-Shia Friends)

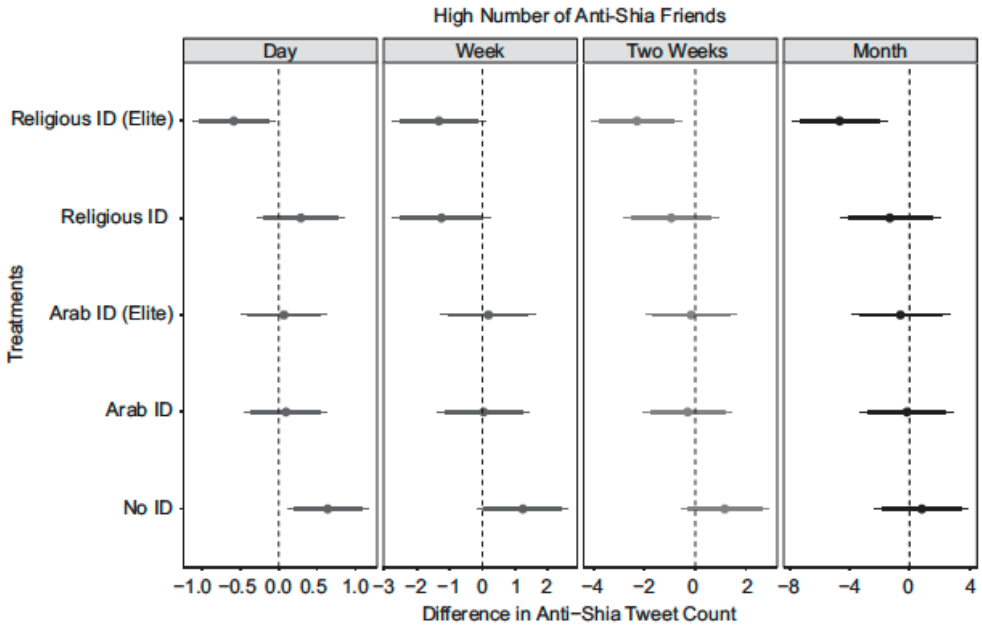


5b: High Number of Anti-Shia Friends

```
#High Anti-Shia Friend Network
month<-lm(month_anti_shia_post-month_anti_shia_pre ~ treatment_num,
          data=data_anti_shia_net_high)

two_weeks<-lm(two_weeks_anti_shia_post-two_weeks_anti_shia_pre ~ treatment_num,
              data=data_anti_shia_net_high)

week<-lm(week_anti_shia_post-week_anti_shia_pre ~ treatment_num,
         data=data_anti_shia_net_high)

day<-lm(tpd_anti_shia_post-tpd_anti_shia_pre ~ treatment_num,
        data=data_anti_shia_net_high)

multiplot(day, week, two_weeks, month,
          coefficients=c("treatment_num1", "treatment_num2",
                         "treatment_num3", "treatment_num4", "treatment_num5"),
          newNames=c(treatment_num1="Arab ID ",
                     treatment_num2="Religious ID ",
                     treatment_num3="Arab ID (Elite)",
                     treatment_num4="Religious ID (Elite)",
                     treatment_num5=" No ID "),
          names=c("     Day", "   Week", "  Two Weeks ", "Month    "),
          title="Figure 5b: Replication", scales="free_x",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen", "black")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(),
                     legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=15))+
  ylab("Treatments") +
  xlab("Difference in Anti-Shia Tweet Count")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```




```
knitr::include_graphics(path="replication_figures/figure_5b_orig.png")
```



Figures Pt. 2: Survey Experiment

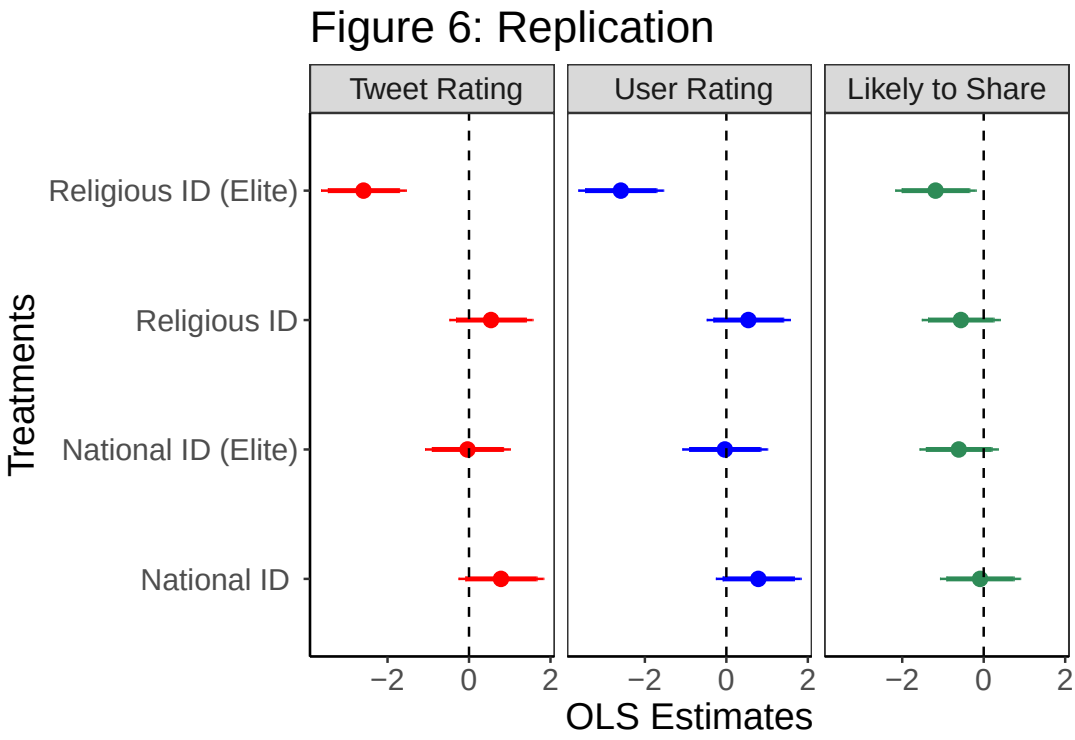
In this part of the report, we replicate the figures from the paper concerning the survey experiment. As in the previous section, we omit the authors' preprocessing code from the report for brevity, but comment on some of their decisions in the section discussing our future work.

Figure 6: Effect of Primes on all Tweet Ratings

While this figure is successfully replicated from the authors' code in the sense that it matches the figure from the paper, the figure in the paper is slightly wrong due to a bug in the code. When fitting "model2" below, the authors use "combined1" as the dependent variable (instead of "combined2"), making the regression the same as "model1". Consequently, the first two columns in the figure, "Tweet Rating" and "User Rating", mistakenly show the exact same coefficients, which are those of the regression of "combined1" on the treatment.

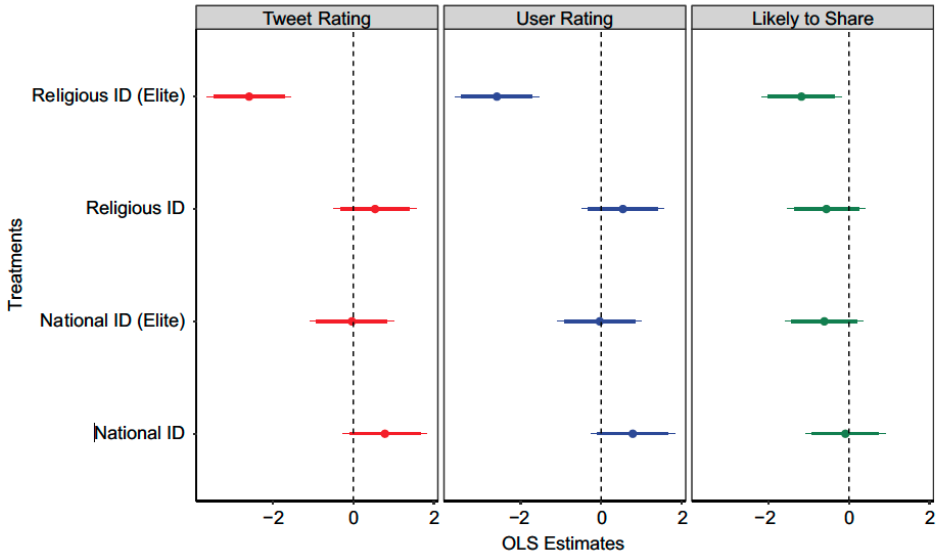
Below, we show the version in the paper, the replication from the authors' code, a corrected version of the figure, and a table comparing the coefficients in the incorrect and correct versions of "model2". Because this snippet of code was copied and pasted for all the plots describing the survey experiment in the authors' code, we fix this error for all subsequent plots in the same way.

```
#OLS Combined Ratings
model1<-lm(combined1~treatment, data=data)
model2<-lm(combined1~treatment, data=data)
model3<-lm(combined3~treatment, data=data)

multiplot(model1, model2, model3,
          coefficients=c("treatment2", "treatment3", "treatment4", "treatment5"),
          newNames=c(treatment2="Religious ID",
                     treatment3="National ID ",
                     treatment4="Religious ID (Elite)",
                     treatment5="National ID (Elite)"),
          names=c(" Tweet Rating", " User Rating", "Likely to Share"),
          title="Figure 6: Replication",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(), legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=16))+
  ylab("Treatments") + xlab("OLS Estimates")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```



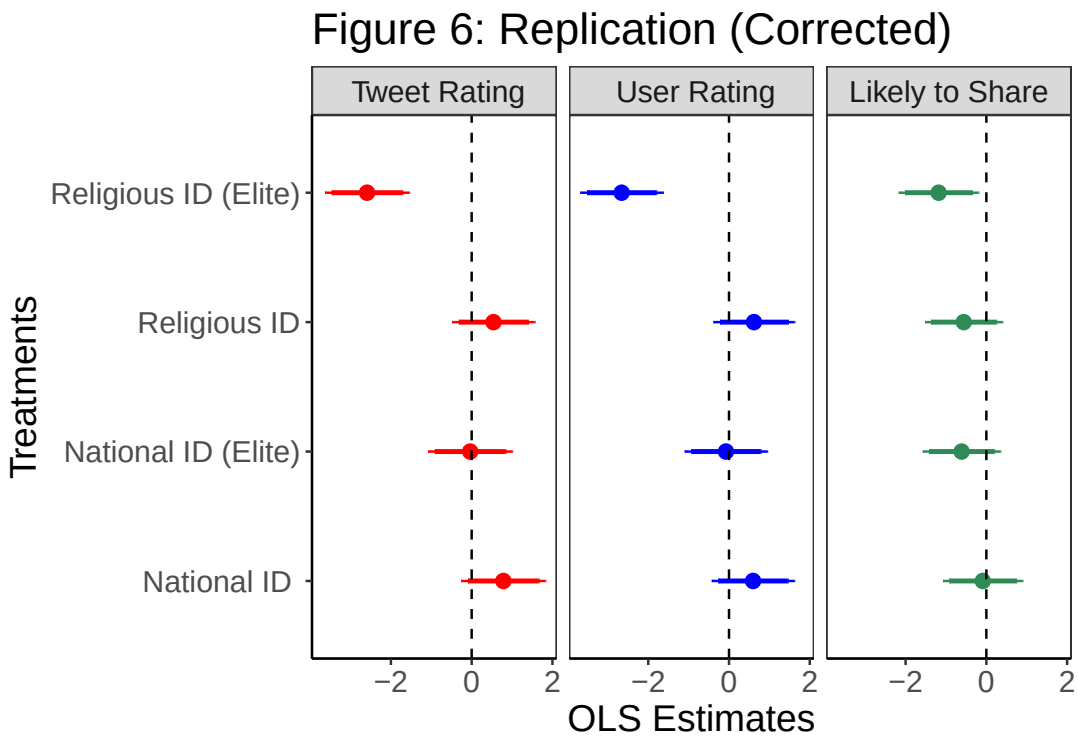

```
knitr::include_graphics(path="replication_figures/figure_6_orig.png")
```

FIGURE 6. Effect of Primes on all Tweet Ratings




```
####FIXED

model2_fixed<-lm(combined2~treatment, data=data)

multiplot(model1, model2_fixed, model3,
          coefficients=c("treatment2", "treatment3", "treatment4", "treatment5"),
          newNames=c(treatment2="Religious ID",
                     treatment3="National ID ",
                     treatment4="Religious ID (Elite)",
```

```
                    treatment5="National ID (Elite)"),
          names=c(" Tweet Rating", " User Rating", "Likely to Share"),
          title="Figure 6: Replication (Corrected)",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(), legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=16))+
  ylab("Treatments") + xlab("OLS Estimates")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```




```
stargazer(model2, model2_fixed,
          column.labels = c("Incorrect DV", "Correct DV"),
          covariate.labels = c(treatment2="Religious ID",
                               treatment3="National ID ",
                               treatment4="Religious ID (Elite)",
                               treatment5="National ID (Elite)"),
          type="latex", header=FALSE,
          title = "Model 2 (User Rating): Incorrect vs Correct DV")
```

Table 1: Model 2 (User Rating): Incorrect vs Correct DV

	Dependent variable:	
	combined1 Incorrect DV	combined2 Correct DV
	(1)	(2)
Religious ID	0.540	0.618
	(0.521)	(0.511)
National ID	0.785	0.595
	(0.531)	(0.521)
Religious ID (Elite)	-2.590^{***}	-2.659^{***}
	(0.531)	(0.521)
National ID (Elite)	-0.037	-0.075
	(0.531)	(0.521)
Constant	-4.099^{***}	-3.909^{***}
	(0.388)	(0.380)
Observations	328	328
R^2	0.147	0.153
Adjusted R^2	0.136	0.142
Residual Std. Error (df = 323)	2.953	2.897
F Statistic (df = 4; 323)	13.866^{***}	14.577^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

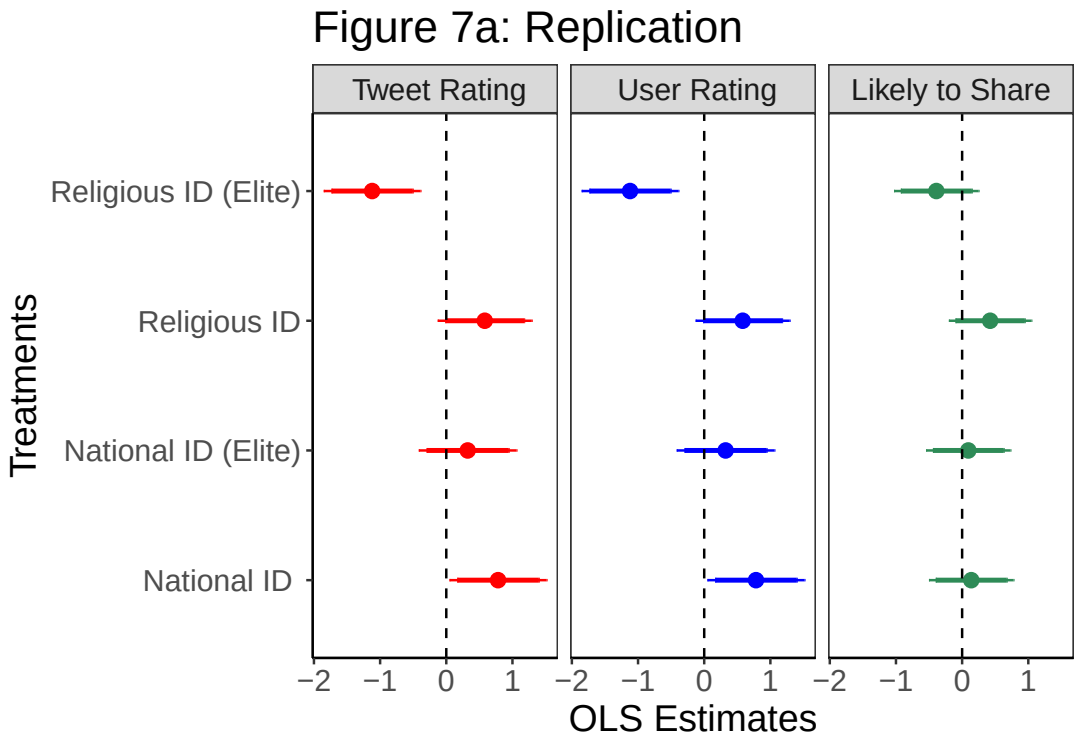
Figure 7: Effect of Primes on (a) Sectarian and (b) Counter-Sectarian Tweet Ratings

Figure 7a:

This figure is successfully replicated from the authors' code, with the aforementioned caveat regarding "Model 2".

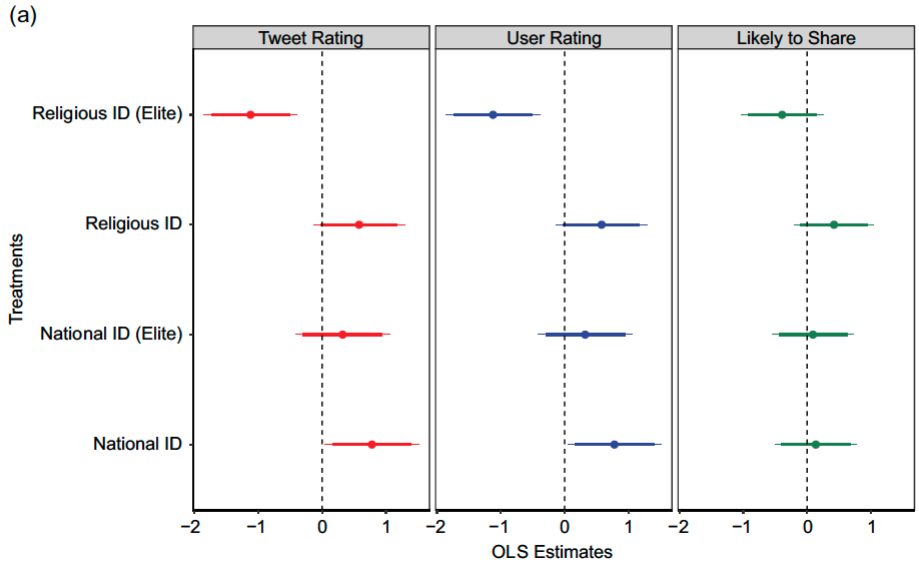
In describing this figure, the authors write "...However, the common-national-identity treatment (without elite support) actually had a backlash effect on sectarian tweet ratings, increasing favorability ratings of both the tweets themselves and the users who sent them." If we test this hypothesis at the 5% level, we would reject the null and agree with this statement for "Model 1" (Tweet Rating), and this is easy to see visually as well. However, we would fail to reject the null hypothesis when it came to "Model 2" (User Rating), but this is only visually apparent once we plot the correct model. We also see this in Table 2 below.

```
#OLS Sectarian Ratings
model1<-lm(sec1~treatment, data=data)
model2<-lm(sec1~treatment, data=data)
model3<-lm(sec3~treatment, data=data)

multiplot(model1, model2, model3,
          coefficients=c("treatment2", "treatment3", "treatment4", "treatment5"),
          newNames=c(treatment2="Religious ID",
                     treatment3="National ID ",
                     treatment4="Religious ID (Elite)",
                     treatment5="National ID (Elite)"),
          names=c(" Tweet Rating", " User Rating", "Likely to Share"),
          title="Figure 7a: Replication",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(), legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=16))+
  ylab("Treatments") + xlab("OLS Estimates")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```



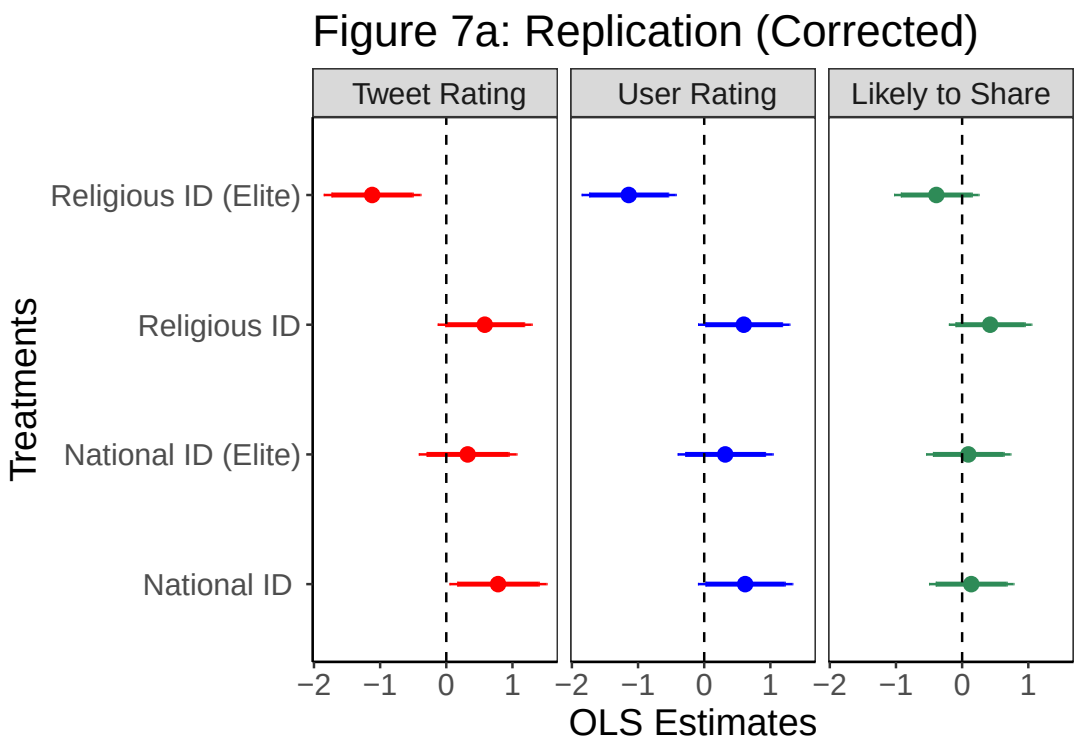

```
knitr::include_graphics(path="replication_figures/figure_7a_orig.png")
```

FIGURE 7. Effect of Primes on (a) Sectarian and (b) Counter-Sectarian Tweet Ratings




```
###FIXED

model2_fixed<-lm(sec2~treatment, data=data)

multiplot(model1, model2_fixed, model3,
          coefficients=c("treatment2", "treatment3", "treatment4", "treatment5"),
          newNames=c(treatment2="Religious ID",
                     treatment3="National ID ",
```

```
                    treatment4="Religious ID (Elite)",
                    treatment5="National ID (Elite)"),
          names=c(" Tweet Rating", " User Rating", "Likely to Share"),
          title="Figure 7a: Replication (Corrected)",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(), legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=16))+
  ylab("Treatments") + xlab("OLS Estimates")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```




```
stargazer(model1, model2_fixed,
          column.labels = c("Model 1", "Model 2 w/ Correct DV"),
          covariate.labels = c(treatment2="Religious ID",
                               treatment3="National ID ",
                               treatment4="Religious ID (Elite)",
                               treatment5="National ID (Elite)"),
          type="latex", header=FALSE,
          title = "Effects on Tweet and User Rating for Figure 7a",
          dep.var.labels = c("Tweet Rating", "User Rating"))
```

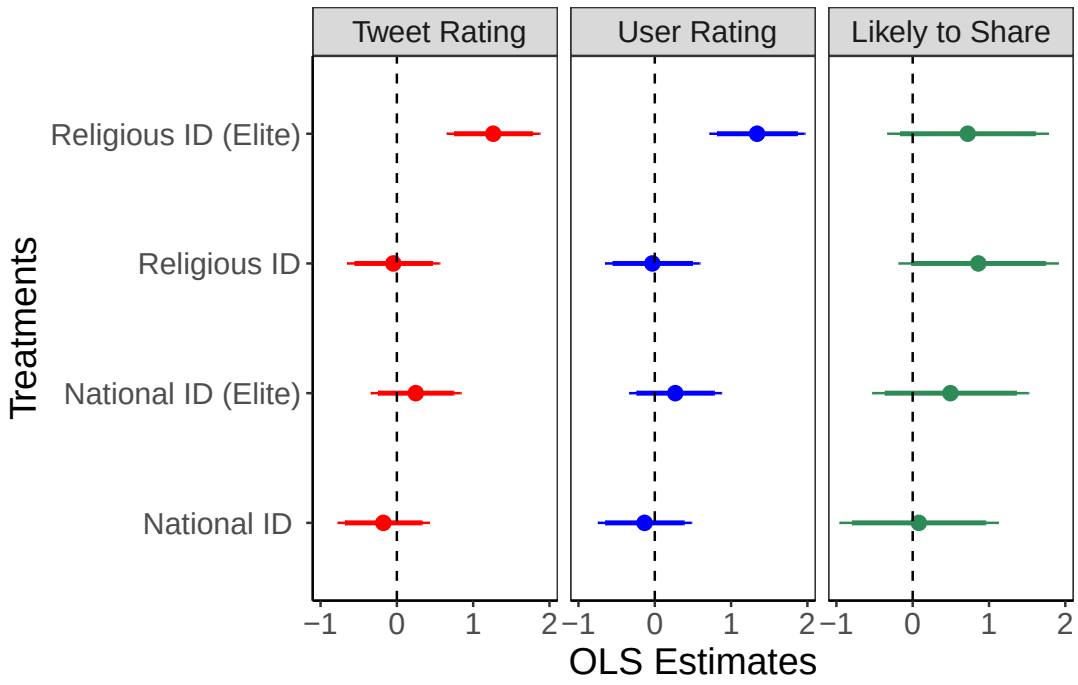
Table 2: Effects on Tweet and User Rating for Figure 7a

	Dependent variable:	
	Tweet Rating Model 1	User Rating Model 2 w/ Correct DV
	(1)	(2)
Religious ID	0.581	0.598*
	(0.363)	(0.353)
National ID	0.783**	0.619*
	(0.374)	(0.364)
Religious ID (Elite)	−1.121***	−1.141***
	(0.373)	(0.363)
National ID (Elite)	0.323	0.318
	(0.377)	(0.367)
Constant	4.395***	4.336***
	(0.272)	(0.265)
Observations	362	362
R^2	0.089	0.089
Adjusted R^2	0.079	0.079
Residual Std. Error (df = 357)	2.179	2.119
F Statistic (df = 4; 357)	8.709***	8.697***
Note:		*p<0.1; **p<0.05; ***p<0.01

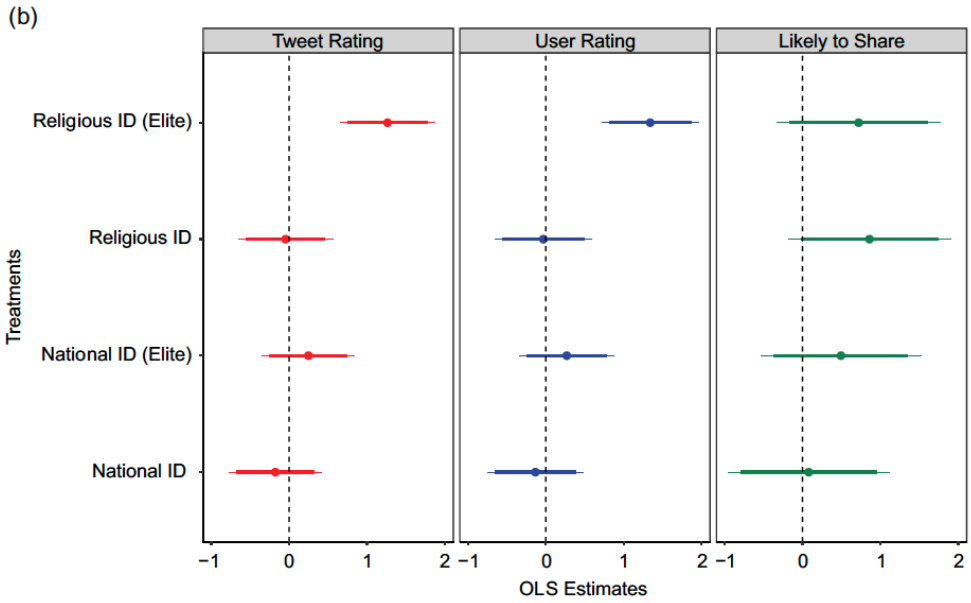
Figure 7b:

This figure is successfully replicated from the authors' code. Oddly enough, they caught the "Model 2" issue with the copy-pasting for this figure, but didn't correct it for the previous figures.

```
#OLS Counter Sectarian Ratings
model1<-lm(counter_sec1~treatment, data=data)
model2<-lm(counter_sec2~treatment, data=data)
model3<-lm(counter_sec3~treatment, data=data)

multiplot(model1, model2, model3,
          coefficients=c("treatment2", "treatment3", "treatment4", "treatment5"),
          newNames=c(treatment2="Religious ID",
                     treatment3="National ID ",
                     treatment4="Religious ID (Elite)",
                     treatment5="National ID (Elite)"),
          names=c(" Tweet Rating", " User Rating", "Likely to Share"), title="",
          sort="alphabetical",
          innerCI=1.645, outerCI=1.96, single=FALSE,
          zeroType = 0,legend.position="none") +
  scale_color_manual(values=c("red", "blue", "seagreen")) +
  theme_bw() + theme(panel.grid.major = element_blank(),
                     panel.grid.minor = element_blank(), legend.position="none",
                     axis.line = element_line(colour = "black"),
                     text = element_text(size=16))+
  ylab("Treatments") + xlab("OLS Estimates")+
  geom_vline(aes(xintercept = 0), size = .5, linetype = "dashed")
```




```
knitr::include_graphics(path="replication_figures/figure_7b_orig.png")
```



Where We're Going Next

We've identified several directions to remedy potential misspecification and extend the analysis, the most notable of which are:

- **Unstable treatment effects:** With the accumulation of "treatment tweets" in the sockpuppet account used for the Twitter experiment, we suspect that the study's treatment would be unstable over time. The cumulative amount of nearly-identical tweets for the last person treated versus the first makes the treatment less effective if a user discounts a reply from an account they assume is a bot. We could investigate this by interacting the treatment with a coarsened variable representing when the subject was treated.
- **Clustered standard errors:** The survey data included information on the subject's geography, which was not used in either of the basic OLS estimates reported in the paper, or the controlled regressions in the appendix. Given the context of Lebanon's consociational democracy and ethnic distribution based on religious sect, we want to cluster the standard errors of the treatment effect by the different geographic/administrative levels provided to assess the robustness of the findings.
- **Item-level variation in survey experiment:** In analyzing the survey data, the authors collapsed each person into a single data point by averaging over all of their measurements (i.e., across all the tweets that the subjects were asked to rate). We could expand on this analysis by exploring item-level variation in the efficacy of the treatments.
- **Exploratory analysis of actual tweets:** We scraped all the tweets from their sockpuppet account, and are analyzing them to see what additional context we can add to the results. For example, not a single "treatment tweet" administered by the account received a single form of engagement (e.g. reply, retweet, favorite, etc.). Social media experimenters we've consulted have suggested that it is unusual to observe an effect of the magnitude described in the paper with such little engagement.