An extension of Siegel & Badaan 2020 (Working title)

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

11/29/2021

Note to the reader

This is a rough draft of our extension of the work in "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online". We split our work into two main branches: one focused on conceptual extensions that address how the data is aggregated and collected, and one focused on methodological extension, in which we explore the effect of different model specifications on the conclusions of the study.

The RMarkdown file to generate this document can be found in our GitHub repo, and contains cells to calculate quantities we discuss in the text, as well as the figures shown.

Conceptual Extension: Background

In the original paper, the authors conduct two experiments to estimate the effect of counter-sectarian primes on sectarian attitudes: one on Twitter, and another through a field survey. In this section, we respond to the survey experiment, highlighting experimental and analytical decisions that we believe obscure important underlying dynamics of the phenomenon of interest, and the context in which it was studied.

In the survey experiment, a firm approached a representative sample of 500 participants from Lebanese society and asked them to rate a total of 8 tweets, 4 of which were "sectarian" and 4 of which were "counter-sectarian" in nature. The participants were asked to rate the content of the tweet itself, the person who wrote the tweet, and how likely they were to share the tweet themselves, all on a scale of 1-10, with lower values indiciating negative perception and lower likelihood to share. Participants were randomly assigned to either a control condition, or to receiving one of four priming statements before rating the tweets (translated from Arabic by the authors):

- **Control statement:** Over the past few years, there has been arise in sectarian tensions in Lebanon and across the Middle East.
- **National identity prime:** Over the past few years, there has been arise in sectarian tensions in Lebanon and across the Middle East. *But many people agree that their sect does not make them better than anyone else. They agree that we are all Lebanese and we all should be equal. We all live on one land. We share the same history and the same future; we share the same culture, the same food, and language. Most importantly, we share a common Lebanese identity.*
- **Religious identity prime:** Over the past few years, there has been arise in sectarian tensions in Lebanon and across the Middle East. *But many people agree that their sect does not make them better than anyone else. They agree that we all believe in one God and we should all be equal. We all live on one land, we share the same history and the same future; we share the same culture, the same food, and language. Most importantly, we share a common belief in God.*
- **Elite-endorsed national ID prime:** " Over the past few years, there has been a rise in sectarian tensions in Lebanon and across the Middle East. *But many prominent politicians including members of the March 8 bloc, the March 14 bloc, and independents have called for people to come together. They agree that we are all Lebanese...*

- **Elite-endorsed religious ID prime:** " Over the past few years, there has been a rise in sectarian tensions in Lebanon and across the Middle East. *But many prominent Christian, Sunni, and Shia religious leaders have issued religious decrees calling for people to come together and stop inciting sectarian hatreds. They agree that we all believe in one God...*

The original paper measures the treatment effects on three outcomes of interest: (1) the average rating across all sectarian tweets; (2) the average rating across all counter-sectarian tweets; and (3) total counter-sectarian sentiment, which they define as the difference between (2) and (1). In this response, we argue that the way the outcomes were aggregated in the original paper obscures important heterogeneity in sentiment and response patterns, and introduces complications in understanding the external validity of the study's results. Furthermore, we discuss elements of the experimental design that cast doubt on how well the collected data represents the intended quantity of interest.

Conceptual Extension: What is the actual/collected QoI here?

In this survey experiment, people are visited in-person by an employee of the agency conducting the survey, and are asked to rate the aforementioned tweets on an iPad. Presumably, they are conscious of the fact that they are being observed by the surveyor – the fact that there is a considerable number of cases where participants refuse to respond to some tweets is evidence of this social pressure. In spirit, the research intends to measure sectarian attitudes and uses ratings provided in the survey as a proxy of such. In practice, what's being measured is *professed attitude to an observer*, which is subtly but importantly different.

The implication is that the estimated treatment effect of the primes is actually how *what a respondent tells you* will change, but not necessarily *what they actually feel*, and there's limited reason to expect the two quantities to align, especially when it comes to contentious sectarian content. Under this design, responses may be performative, with subjects answering the way they *think* the surveyor wants them to, or in a way deemed "socially acceptable". Given this tendency, it's difficult to understand the practical value of these interventions, and it would be interesting to explore techniques that can help minimize the gap between the two concepts by abstracting the situation away from the self.

Another peculiarity is that the ratings of the tweets and users closely mirror each other. Across all tweets, ~75% of the ratings match, and ~90% have a difference ≤ 1. This raises the question of whether tweet and user favorability truly are actually two different but tightly coupled concepts, or whether there may have been something in the phrasing of the survey that lead to this collapse of the two concepts into one.

Conceptual Extension: What is actually estimated here?

As confirmed in communication with the authors, survey subjects occasionally refused to answer some of the evaluations, and these were recorded in the data as null values. Out of 500 participants, 28% refused to answer at least one query about a sectarian tweet, and 20% refused to answer at least one query about a counter-sectarian tweet, with 12% of participants refusing to answer at least one query in both categories. Consequently, when the original paper aggregates across tweets by taking an average, any participant that refused to evaluate any of the relevant tweets would have a null value for their aggregated measure. This not only dramatically shrinks the sample size, but also invalidates the work that the expensive surveying firm was hired to do: the sample is no longer representative of Lebanon as a whole, but of those that agreed to evaluate each and every tweet, which is a sub-sample that may differ considerably and is difficult to understand. Because of this, it is difficult to interpret the estimated treatment effects as an average treatment effect for a random Lebanese citizen.

Conceptual Extension: Disaggregation exposes a more complex sectarian-political dynamic

The paper reports that the each of the sectarian tweets targets a single sect, with the targeted sects being Sunnis, Shia, Christians, and Druze. On the other hand, the counter-sectarian tweets are generically counter-sectarian, i.e. they call for unity, but do not promote a specific sect. Because people from different sects may respond differently to sectarian tweets and counter-sectarian primes depending on the relationship between their personal sect and the targeted sect, it would be reasonable to expect different treatment efficacy for each of the sectarian tweets, which we explore in this section.

To contextualize the differences observed between individual tweets in the survey experiment, we contacted the authors for the exact phrasing of the tweets used in Arabic. As native Arabic speakers, we are able to faithfully translate these into English below, and discuss how the nuance of the phrasing affects this experiment:

Sectarian tweets:

1. Now the Arab people (or most of them) are certain that *Hezb Al-lat* is a barbaric, destructive arm of Iran in the region and that it supports its *rawafid* followers #HezbAllah_is_Lebanons_Destruction
2. I am a muslim [translator note: female] and I hate non-believers because they do not worship God [Allah].
3. To the Christians of the March 8 bloc, how can you allow your future generations to live like you, as followers of an armed party that assassinated the best men of Lebanon? #Lebanon needs a new generation
4. #Shia muslims spend the night praying to God [Allah] on *Laylat Ul-Qadr* [translator note: one of the holiest nights of the year in Islam], while *Da'esh* [ISIS] scum blow up a masjid they are praying in. *Hasbuna Allah Wa Neam Alwakeel* [translator note: sect-neutral religious saying along the lines of "We leave this to God, and he is the best agent."] #Kabul

Counter-sectarian tweets:

1. Did you know that, in Lebanon, we are all Shia if a Shia is threatened, we are all Sunni if a Sunni is threatened, and we are all Christians if a Christian is threatened? #WeAreAllLebanon
2. There is no hope for Lebanon as long as politicians keep speaking in sectarian logic: Sunni, Shia, Druze, Christian, but nobody says "I am Lebanese". Sectarianism is cancer.
3. To the devoted sons of the Sunni sect, to the devoted sons of the Shia sect, we do not blame you for a lowly, divisive minority that does not represent you #NoToSectarianism
4. Lebanon has no economy, no justice system, no oversight, no electoral law, no establishments, no leadership, no authority... etc. Sectarianism infiltrates all of these headlines. #NoToSectarianism

First, we note that contrary to the paper's claim, none of the sectarian tweets refer to Druze specifically. Second, it's important to note that, with the exception of #2, each of these sectarian tweets is confounded by considerable geopolitical undertones: #1 refers to Iran, #3 is more readily interpreted as an attack on either of the "Free Patriotic Movement" or "Marada" parties, and #4 mixes in the complexities of ISIS perception. In a country where party and religious lines nearly coincide, it is difficult to disentangle political and sectarian sentiment.

Sectarian tweet #2 highlights a flaw in the experimental design; although it promotes a form of hate, that hate is more accurately described as "anti-atheist" rather than sectarian, as it describes hatred for those who do not believe in or worship God. All the primes emphasize shared identity, with the religious primes specifically highlighting a shared belief in God among religions, and this is likely why we observe in Figure 1 that the primes actually *increase* support for this tweet (with the exception of the religious elite prime, which has no effect). This example alone highlights the heterogeneity in item-level effects and potential unintended harms, which is enough to warrant the disaggregation of the analysis. Furthermore, by disaggregating, we

can argue against the following claim in the paper: "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". We see that, in fact, the national ID prime had null effects on all sectarian tweets except tweet #2, which we've demonstrated is categorically different from the rest; the purported "backlash effect" is not supported by the data.

Interestingly, when comparing Figure 1 to Figure 2, we see that treatment effects are more stable across individual tweets in the counter-sectarian case. This is in line with our earlier prediction that treatment effects would differ for the sectarian tweets depending on the sect; when the tweets do not promote/hate any specific sect, the treatment effects should be less distinguishable from each other.

Fig 1. Effect of treatments on sectarian tweet evaluations

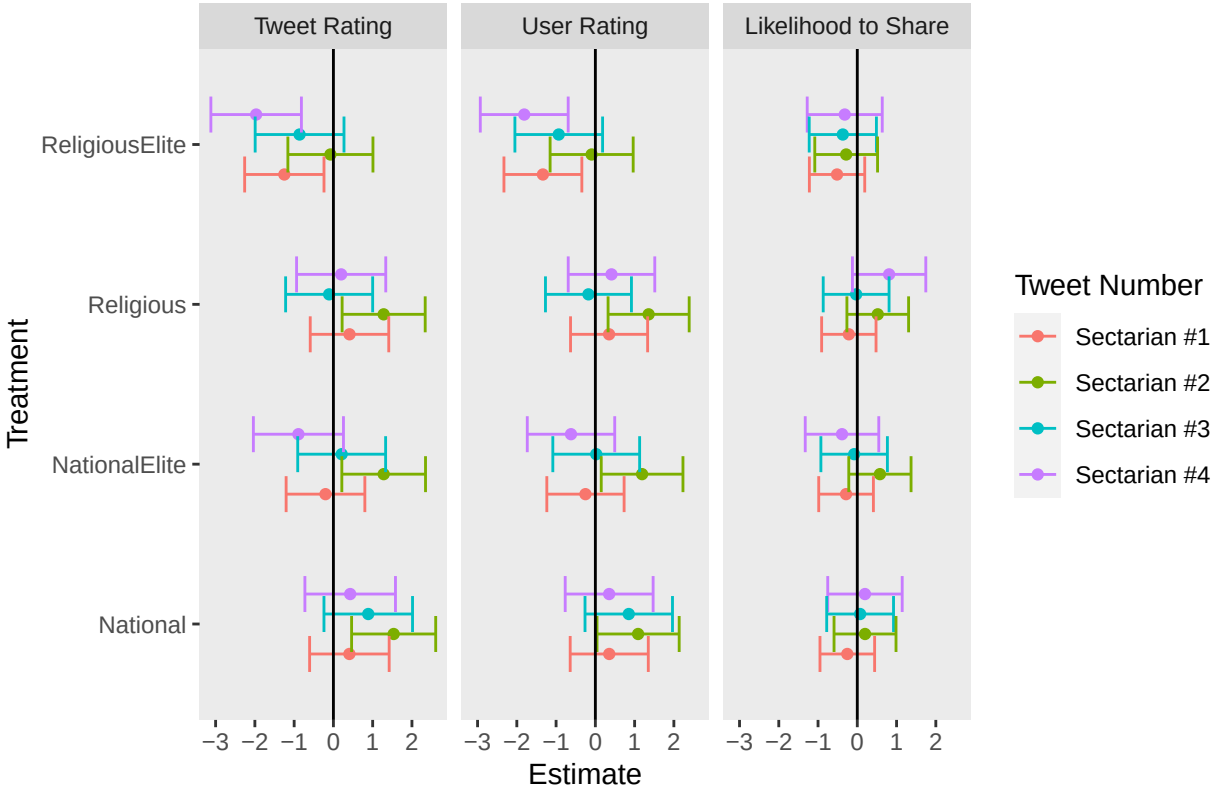
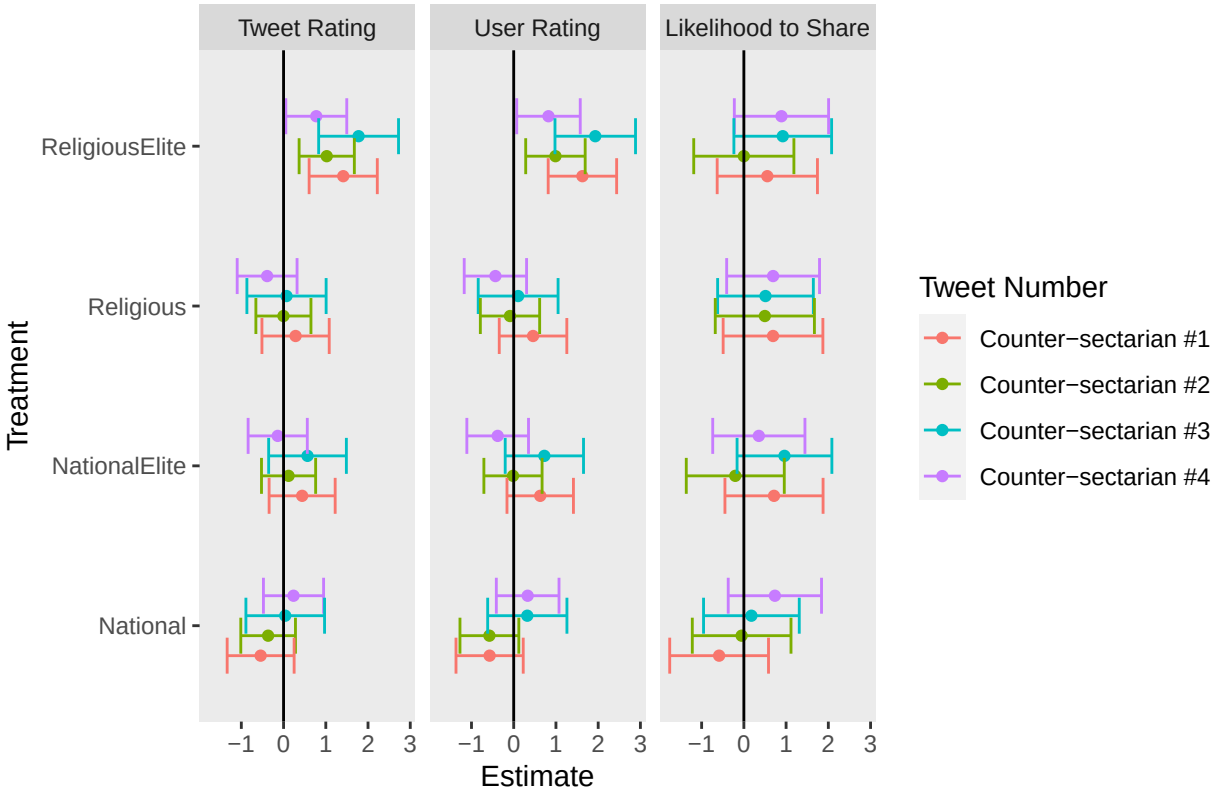


Fig 2. Effect of treatments on sectarian tweet evaluations



Conceptual Extension: Exploring patterns of non-response

The original paper reports that "...one of the strongest predictors of unfavorable ratings for sectarian tweets and favorable ratings for counter-sectarian tweets in our survey experiment was users' level of motivation to control prejudice (MCP)–their concern with acting prejudiced or being perceived as prejudiced." To measure this, they regress the aggregated composite measure highlighted in the background section of this report on several controls, including MCP. Putting aside our concerns regarding the aggregation, the patterns of non-response present an interesting opportunity to test the validity of the MCP measure. Following the logic outlined by the authors, MCP should be predictive of non-response, as fear of impropriety is supposedly the motive behind refusing to answer. In Tables 1 & 2, we use the vast set of controls employed by the original paper in their robustness tests, and find that MCP is not meaningfully associated with non-response for both sectarian and counter-sectarian tweets. This could either cast doubt on the validity of MCP as a measure, or suggest that the underlying reason for non-response is not what we expect it to be.

On the other hand, and quite intuitively, we find that the respondent's sect has a large and significant effect on non-response for sectarian tweets (with the exception of #2, which we have previously demonstrated is categorically different). We find no reliable predictors of non-response for counter-sectarian tweets, making a puzzling behavior even more puzzling.

Table 1: Estimating effects on non-response for sectarian tweets

	Dependent variable:			
	NA Tweet 1	NA Tweet 2	NA Tweet 3	NA Tweet 4
	(1)	(2)	(3)	(4)
Christian (ref: Shia)	0.144***	0.053	0.087**	0.069*
	(0.041)	(0.042)	(0.041)	(0.041)
Druze	0.071	−0.066	−0.0004	−0.011
	(0.065)	(0.067)	(0.066)	(0.065)
Sunni	0.098**	0.020	0.113**	0.095**
	(0.044)	(0.045)	(0.044)	(0.044)
Political Interest	0.014	0.011	0.008	0.012
	(0.012)	(0.012)	(0.012)	(0.012)
Social Media Use	0.017	0.034***	0.025*	0.022*
	(0.013)	(0.013)	(0.013)	(0.013)
Religiosity	0.096*	0.098*	0.014	0.025
	(0.057)	(0.059)	(0.058)	(0.057)
Sectarianism Index	0.050**	−0.034	−0.015	0.015
	(0.023)	(0.024)	(0.023)	(0.023)
Sectarian System Justification Index	−0.030	−0.017	−0.008	0.018
	(0.029)	(0.030)	(0.029)	(0.029)
MCP	0.017	−0.024	0.031	0.004
	(0.033)	(0.033)	(0.033)	(0.033)
Constant	−0.596***	−0.109	−0.158	−0.278
	(0.227)	(0.233)	(0.230)	(0.227)
Observations	372	372	372	372
R^2	0.070	0.047	0.049	0.042
Adjusted R^2	0.047	0.023	0.026	0.018
Residual Std. Error (df = 362)	0.278	0.285	0.281	0.278
F Statistic (df = 9; 362)	3.036***	1.969**	2.092**	1.771*
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: Estimating effects on non-response for counter-sectarian tweets

	Dependent variable:			
	NA Tweet 1	NA Tweet 2	NA Tweet 3	NA Tweet 4
	(1)	(2)	(3)	(4)
Christian (ref: Shia)	0.025	0.011	0.054*	0.007
	(0.021)	(0.023)	(0.030)	(0.023)
Druze	−0.013	−0.029	−0.019	−0.016
	(0.034)	(0.036)	(0.047)	(0.036)
Sunni	−0.018	−0.020	0.025	0.017
	(0.023)	(0.024)	(0.032)	(0.024)
Political Interest	0.005	0.004	0.013	0.011
	(0.006)	(0.007)	(0.009)	(0.007)
Social Media Use	0.005	0.012*	0.005	0.002
	(0.007)	(0.007)	(0.009)	(0.007)
Religiosity	−0.023	0.009	−0.087**	0.024
	(0.030)	(0.032)	(0.041)	(0.032)
Sectarianism Index	−0.002	−0.005	0.024	−0.006
	(0.012)	(0.013)	(0.017)	(0.013)
Sectarian System Justification Index	−0.001	0.009	−0.005	−0.006
	(0.015)	(0.016)	(0.021)	(0.016)
MCP	0.004	−0.019	−0.013	−0.011
	(0.017)	(0.018)	(0.024)	(0.018)
Constant	0.060	0.017	0.230	−0.016
	(0.119)	(0.126)	(0.165)	(0.126)
Observations	372	372	372	372
R^2	0.025	0.023	0.040	0.017
Adjusted R^2	0.001	−0.001	0.016	−0.008
Residual Std. Error (df = 362)	0.145	0.154	0.201	0.154
F Statistic (df = 9; 362)	1.043	0.944	1.691*	0.687
Note:				*p<0.1; **p<0.05; ***p<0.01

Methodological Extension #1: Potentially unstable treatment effects

In the Twitter experiment conducted by the authors, they use a "sockpuppet" Twitter account, which is a fake persona designed to look like a real user. The experiment is to reply to 957 random accounts tweeting sectarian content with one of the 4 primes used in the survey experiment (albeit with slightly different phrasing), and then measure whether this intervention decreases their production of sectarian tweets.

Because the replies coming from the sockpuppet account are identical, if a subject treated later in the experiment opened the sockpuppet's account page and saw what appeared to be "spammy" replies (hundreds of replies with exactly the same phrasing), the subject might discount the sockpuppet as a bot and the treatment may lose efficacy.

To explore this potential effect, we split the data into 3 groups based on when they were treated: the first third, the second third, and the last third. We then measure the interaction between the treatment and this coarsened order of treatment, and find no substantial evidence that this dynamic is occurring, as indicated by the null interaction effects in Table 3.

Methodological Extension #2: clustered standard errors over geographical area

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. The nature of Lebanon's consociational democracy today consists of ethnic and sectarian cleavages and enhanced polarization. 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.

The three geographical data points collected in the survey were Mohafaza, Caza, and City. Since "City" is nested within the other two variables (as it is the smallest of the terms), we chose to cluster by City the models included in the survey data analysis. These include the effect of the primes on all tweet ratings as well as the effect of the primes on sectarian and counter-sectarian tweet ratings. We find that the coefficients are the same with only slight differences in standard errors. This indicates that the conclusions drawn do not deviate even when clustering by location, as shown in Table 4 (for the combined rating, as described in the background), Table 5 (for the sectarian tweets), and Table 6 (for the counter-sectarian tweets). We use the aggregated measure from the original paper as the dependent variable here to hold that part constant in comparison with the paper, such that the only difference here is in how the standard errors are clustered.

Table 3: Interaction between Twitter intervention and intervention group

	Dependent variable:			
	Day	Week	2 Weeks	Month
	(1)	(2)	(3)	(4)
Arab ID	−0.025	−0.271	−0.673	−0.057
	(0.378)	(0.936)	(1.305)	(2.914)
Religious ID	−0.245	0.132	0.857	7.141**
	(0.401)	(0.992)	(1.378)	(3.125)
Arab ID (Elite)	0.635	0.222	−1.373	−0.715
	(0.393)	(0.973)	(1.363)	(3.075)
Religious ID (Elite)	−1.772***	−1.165	−3.727***	−5.650*
	(0.399)	(0.987)	(1.378)	(3.143)
No ID	0.478	0.840	0.556	−0.437
	(0.386)	(0.959)	(1.326)	(3.060)
Group 2	−0.231	−0.373	−0.206	2.200
	(0.393)	(0.977)	(1.378)	(3.181)
Group 3	0.218	−0.037	0.051	1.393
	(0.465)	(1.148)	(1.651)	(3.723)
Arab ID* Group2	−0.122	0.432	0.486	−1.197
	(0.548)	(1.357)	(1.902)	(4.225)
Religious ID *Group2	0.148	−0.265	−1.645	−8.219*
	(0.589)	(1.456)	(2.036)	(4.539)
Arab ID (Elite)*Group2	−0.571	−0.093	1.417	0.423
	(0.552)	(1.372)	(1.934)	(4.384)
Religious ID (Elite)*Group2	1.000*	0.216	1.824	−0.200
	(0.556)	(1.379)	(1.939)	(4.432)
No ID*Group2	−1.032*	−1.780	−2.316	−4.974
	(0.536)	(1.331)	(1.863)	(4.240)
Arab ID* Group3	0.249	0.667	0.759	0.238
	(0.592)	(1.462)	(2.088)	(4.628)
Religious ID *Group3	0.330	−1.267	−1.686	−9.142*
	(0.605)	(1.491)	(2.121)	(4.782)
Arab ID (Elite)*Group3	−0.699	−0.405	0.970	0.382
	(0.609)	(1.501)	(2.154)	(4.856)
Religious ID (Elite)*Group3	1.225**	−1.503	0.778	−2.026
	(0.605)	(1.492)	(2.121)	(4.775)
No ID*Group3	−0.737	−1.050	−1.600	1.798
	(0.595)	(1.474)	(2.091)	(4.740)
Constant	0.058	0.451	1.373	0.650
	(0.278)	(0.691)	(0.959)	(2.249)
Observations	952	944	922	795
R^2	0.074	0.028	0.029	0.046
Adjusted R^2	0.057	0.010	0.010	0.026
Residual Std. Error	2.005 (df = 934)	4.935 (df = 926)	6.850 (df = 904)	14.225 (df = 777)
F Statistic	4.392*** (df = 17; 934)	1.550* (df = 17; 926)	1.572* (df = 17; 904)	2.226*** (df = 17; 777)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Comparing regular and clustered SEs for estimates of treatment effect on combined tweet rating

	Dependent variable:					
	Tweet Rating		User Rating		Likelihood to Share	
	Regular SE	Clustered	Regular SE	Clustered	Regular SE	Clustered
	(1)	(2)	(3)	(4)	(5)	(6)
Religious ID	0.540	0.540	0.618	0.618	−0.560	−0.560
	(0.521)	(0.472)	(0.511)	(0.435)	(0.489)	(0.519)
National ID	0.785	0.785	0.595	0.595	−0.088	−0.088
	(0.531)	(0.621)	(0.521)	(0.597)	(0.503)	(0.575)
Religious ID (Elite)	−2.590***	−2.590***	−2.659***	−2.659***	−1.182**	−1.182*
	(0.531)	(0.552)	(0.521)	(0.526)	(0.504)	(0.642)
National ID (Elite)	−0.037	−0.037	−0.075	−0.075	−0.612	−0.612
	(0.531)	(0.454)	(0.521)	(0.448)	(0.490)	(0.460)
Constant	−4.099***	−4.099***	−3.909***	−3.909***	−1.888***	−1.888***
	(0.388)	(0.455)	(0.380)	(0.416)	(0.360)	(0.419)
Observations	328		328		377	
R^2	0.147		0.153		0.019	
Adjusted R^2	0.136		0.142		0.008	
Residual Std. Error	2.953 (df = 323)		2.897 (df = 323)		2.994 (df = 372)	
F Statistic	13.866*** (df = 4; 323)		14.577*** (df = 4; 323)		1.796 (df = 4; 372)	
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 5: Comparing regular and clustered SEs for estimates of treatment effect on sectarian tweet rating

	Dependent variable:					
	Tweet Rating		User Rating		Likelihood to Share	
	Regular SE	Clustered	Regular SE	Clustered	Regular SE	Clustered
	(1)	(2)	(3)	(4)	(5)	(6)
Religious ID	0.581	0.581	0.598*	0.598*	0.424	0.424
	(0.363)	(0.373)	(0.353)	(0.341)	(0.318)	(0.312)
National ID	0.783**	0.783**	0.619*	0.619*	0.138	0.138
	(0.374)	(0.345)	(0.364)	(0.347)	(0.326)	(0.442)
Religious ID (Elite)	−1.121***	−1.121***	−1.141***	−1.141***	−0.390	−0.390
	(0.373)	(0.312)	(0.363)	(0.292)	(0.326)	(0.251)
National ID (Elite)	0.323	0.323	0.318	0.318	0.094	0.094
	(0.377)	(0.251)	(0.367)	(0.254)	(0.325)	(0.272)
Constant	4.395***	4.395***	4.336***	4.336***	2.507***	2.507***
	(0.272)	(0.274)	(0.265)	(0.259)	(0.237)	(0.255)
Observations	362		362		406	
R^2	0.089		0.089		0.018	
Adjusted R^2	0.079		0.079		0.008	
Residual Std. Error	2.179 (df = 357)		2.119 (df = 357)		2.013 (df = 401)	
F Statistic	8.709*** (df = 4; 357)		8.697*** (df = 4; 357)		1.796 (df = 4; 401)	
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 6: Comparing regular and clustered SEs for estimates of treatment effect on counter-sectarian tweet rating

	Dependent variable:					
	Tweet Rating		User Rating		Likelihood to Share	
	Regular SE	Clustered	Regular SE	Clustered	Regular SE	Clustered
	(1)	(2)	(3)	(4)	(5)	(6)
Religious ID	−0.047	−0.047	−0.032	−0.032	0.860	0.860
	(0.308)	(0.366)	(0.316)	(0.380)	(0.533)	(0.640)
National ID	−0.177	−0.177	−0.134	−0.134	0.080	0.080
	(0.305)	(0.344)	(0.312)	(0.340)	(0.530)	(0.618)
Religious ID (Elite)	1.263***	1.263***	1.340***	1.340***	0.721	0.721
	(0.310)	(0.377)	(0.317)	(0.397)	(0.537)	(0.674)
National ID (Elite)	0.248	0.248	0.269	0.269	0.494	0.494
	(0.300)	(0.258)	(0.307)	(0.246)	(0.522)	(0.599)
Constant	8.666***	8.666***	8.419***	8.419***	4.577***	4.577***
	(0.224)	(0.324)	(0.229)	(0.341)	(0.387)	(0.531)
Observations	414		414		433	
R^2	0.067		0.070		0.010	
Adjusted R^2	0.058		0.060		0.0004	
Residual Std. Error	1.924 (df = 409)		1.968 (df = 409)		3.416 (df = 428)	
F Statistic	7.401*** (df = 4; 409)		7.646*** (df = 4; 409)		1.046 (df = 4; 428)	
Note:						*p<0.1; **p<0.05; ***p<0.01