British Journal of Political Science (2022), **52**, 1490–1501
doi:10.1017/S0007123421000582

British Journal of Political Science

RESPONSE

Islamic State's Terrorist Attacks Disengage Their Supporters: Robust Evidence from Twitter

Joan Barceló[1]* and Elena Labzina[2]

[1]New York University-Abu Dhabi, Abu Dhabi, United Arab Emirates and [2]Microsoft, Commercial Software Engineering, Zurich, Switzerland
*Corresponding author. Email: joan.barcelo@nyu.edu

(Received 12 June 2021; revised 26 July 2021; accepted 19 August 2021; first published online 31 January 2022)

Abstract

This article responds to Hansen's (2022) comment on the use of social media data to evaluate the effects of terrorist attacks on related online behavior. Hansen casts doubts on our previous finding that terrorist attacks disengage supporters of terrorist groups. The author speculates that this result *might* not hold when considering the bias introduced by the timing of Twitter account suspensions. We contend that this critique is deficient in several respects. First, Hansen speculates about a possible bias, yet no empirical evidence is offered to support this claim. Second, the author largely ignores our discussion of the role of Twitter account suspensions and our measures of Twitter suspension efforts and Anonymous reporting activity to alleviate this concern. In this article, rather than engaging in a qualitative debate on the amount of bias, we make two contributions. First, we offer an empirical investigation of the fragility of our previous finding to the presence of omitted variable bias. Once we account for Twitter suspension activities, we find that an extreme, unlikely amount of confounding is required to alter the estimated effect of terrorist attacks on disengagement. Second, we employ sequential g-estimation to calculate the average controlled direct effect of terrorist attacks on disengagement after controlling for two intermediate confounders: Twitter suspension behavior and Anonymous reporting activities. The estimated average controlled direct effect indicates that Islamic State's terrorist attacks significantly reduced followers in Islamic State-related Twitter accounts after de-mediating this effect from Twitter suspension efforts and Anonymous reporting activities. Further, we show that this average controlled direct effect is robust to a massive, and implausible, violation of the sequential unconfoundedness assumption. Overall, these analyses show that the timing of Twitter account suspensions, as well as any other confounder, is extremely unlikely to alter our conclusion: Islamic State's terrorist attacks disengage their supporters. We conclude this article by offering guidance on how to address practical challenges in political science research using social media data.

Keywords: terrorism; Islamic State; Twitter; social media

In our contribution "Do Islamic State's deadly attacks disengage, deter, or mobilize supporters?" (Barceló and Labzina 2020), we investigated the impact of terrorist attacks on the number of supporters of terrorist organizations, as quantified by shifts in social media followers. By exploiting the exogenous timing of Islamic State (IS) terrorist attacks, we showed that IS-related attacks reduce the number of followers of IS-related Twitter accounts. We further investigated whether this demobilization effect arises from deterrence or disengagement, showing support for the latter. We then asserted that terrorist attacks demobilize support for terrorist organizations through a process of *disengagement*. Empirically, we analyzed a unique panel dataset of over 13,300 Twitter accounts linked to IS that were tracked throughout a 127-day period in 2017.

Our main results relied on model specifications that include: (1) user-account fixed effects; (2) time fixed effects; and (3) time-varying control variables. The number of accounts suspended by Twitter and the number of accounts reported by Anonymous are the time-varying control variables we examined. The rationale for this model specification is discussed at length in our contribution but will be reiterated here for clarity.

User-account fixed effects are fundamental to our identification strategy for several reasons. They remove important observed and unobserved heterogeneity across accounts—language, geographic location of the user, and demographic characteristics—all of which might affect the number of followers, as well as the likelihood of account suspension. Most importantly, user-account fixed effects minimize "selection bias" in the sample before and after an attack (Barceló and Labzina 2020, 1547). This selection bias may arise "because of shifts in either the type of accounts that Anonymous targets, which may be different in a period before and after an attack, or shifts in the behavior of users, including the opening or closing of accounts" (Barceló and Labzina 2020, 1547). As fixed-effects models do not pool information across different accounts over time, they allow us to compare the same accounts before and after the attacks. By design, they are unaffected by the removal and resurgence of accounts over time (for example, replication bias).

Still, we fully acknowledged that changes in "the time-variant aggressive account suspension efforts by Twitter and the 'Anonymous' hacktivist group in the aftermath of terror attacks" can cause some confounding (Barceló and Labzina 2020, 1548). Consequently, we went to a significant effort to construct two variables that measure changes in the preattack and postattack behavior of both Twitter and Anonymous. For this, first, we created a variable that captures the total number of reports Anonymous sends to Twitter every day. Secondly, we generated a measure of behavioral changes of Twitter that captures the total number of IS-related accounts suspended. Our models with and without these controls supported our substantive conclusion.

In her comment, Hansen (2022) argues that Twitter suspension activities could still be a confounder in our models through four channels: (1) follower loss from suspended accounts; (2) heterogeneous Twitter suspension efforts based on account characteristics; (3) replication bias due to the resurgence of suspended accounts; and (4) replatforming to other social media. However, not all four channels are relevant to our empirical design. Contradictorily, Hansen rules out the second and third pathways deeper in her comment. More specifically, she acknowledges that replication bias "is only problematic in random effects models," not in the fixed-effects models, "as fixed effects models measure effects at a within-unit level" (Hansen 2022, fn. 5, page 7). As the fixed-effects models only exploit differences in followers over time within accounts, time-invariant account characteristics observable to Twitter are naturally controlled for. Therefore, the first and fourth pathways are the only channels of bias worth our attention.

On the whole, Hansen's comment offers a speculative, qualitative piece about plausible ways in which Twitter suspension efforts may confound the relationship between terrorist attacks and the number of followers in IS-related Twitter accounts. Despite our significant efforts to measure and control for Anonymous and Twitter suspension efforts, Hansen still considers our controls insufficient because some unmeasured spillover effects and measurement error may still remain in the residual variance, which could affect our conclusions.

We counter that the presence of some confounding bias is not sufficient to invalidate our estimates. In order to alter our conclusions, the amount of this bias would have to be sufficiently large and in the opposite direction of our effects. Hansen's critique falls short of providing any empirical evidence of whether, how, and by how much our estimates could be affected by Twitter suspension efforts. After accounting for Twitter suspension efforts and Anonymous reporting, it remains unclear what these other confounding variables might be. Secondly, even in the presence of some bias, it remains unclear how strongly correlated it might be with both the timing of attacks and the outcome. Again, we do not claim that these do not exist, but claim that we do not know—and Hansen's comment provides no additional direction to identify the relevant quantities of interest.

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

Rather than engaging in a speculative debate on the amount of confounding bias, we believe that it is more productive to empirically investigate the fragility of the main estimates to deviations of the zero-bias assumption. To empirically examine this, we employ sensitivity analysis to evaluate the amount of omitted variable bias (OVB) that would be required to alter our main findings. While the observed number of suspended Twitter accounts might not perfectly capture all Twitter suspension activity efforts, we argue that it does so to a fair extent. Therefore, we benchmark the amount of confounding bias required to alter our conclusions to the observed number of suspended Twitter accounts.

Our sensitivity analysis reveals a critical result. We find that a confounder *six* and *eleven* times as *strong* as that of the observed number of suspended Twitter accounts would be required to reduce the estimated effect of the number of victims in and outside Europe or the United States, respectively, on the number of IS-related Twitter followers. Consequently, we conclude that the amounts of bias required to remove our estimated effects are implausibly large.

Barceló and Labzina's (2020) original robustness checks, the sensitivity analysis to OVB offered in this article, and Hansen's criticisms all assume that Twitter suspension efforts constitute a confounder in the relationship between attacks and Twitter-follower loss. However, we could consider Twitter suspension efforts an intermediate factor or an alternative mechanism of this relationship. In our original contribution (Barceló and Labzina 2020, 1548), we already cautioned that the behavior of Twitter and/or Anonymous in the aftermath of terrorist attacks could plausibly be posttreatment. As Acharya, Blackwell, and Sen (2016: 514) suggest: "conditioning on a posttreatment variable changes the quantity of interest from an overall average treatment effect to a direct effect of the treatment net the posttreatment variable," inducing posttreatment bias.

In this article, we circumvent this issue by employing sequential g-estimation to calculate the average controlled direct effect (ACDE) of terrorist attacks on the number of followers net of the effect of the intermediate pathway of Twitter suspension activity. The results reveal that the estimated direct treatment effects remain substantively unaltered and statistically significant at conventional levels after controlling for the posttreatment confounder. Further, we explore the sensitivity of the results to violations of the sequential unconfoundedness assumption. We find that the unobserved correlation between Twitter suspension activities and outcome error should be *twenty-six times* as *strong* as the observed correlation between Twitter suspension activities and the outcome variable. This indicates that a massive, implausible amount of bias is required to alter the disengagement effect of terrorist attacks.

Lastly, we evaluate the impact of replatforming on our estimates. Even though our estimates are unlikely to be affected by Twitter suspension efforts, we discuss how current evidence is inconsistent with the logic of replatforming. Finally, we conclude by providing a roadmap for scholars who could address the potential perils of using social media data in conflict research.

Sensitivity to Omitted Variable Bias

In this section, we proceed with a formal sensitivity analysis of plausible confounding bias. As an alternative to Hansen's qualitative debate, we can more usefully discuss "how strong confounding would need to be to substantively alter our conclusions" through sensitivity analysis. For this, we employ the set of recently published sensitivity analysis tools offered by Cinelli and Hazlett (2020).[1] We adopt a bounding approach by benchmarking the sensitivity results to the observable variable that most closely captures Twitter suspension activity, quantified as the daily number of accounts suspended by Twitter.[2] This method allows us to make statements about how unmeasured confounding compares to this observable indicator.

[1]This set of tools can be easily implemented using the R package Sensemakr (Cinelli, Ferwerda, and Hazlett 2020).

[2]It should be noted that all reported estimates in Barceló and Labzina's (2020) original contribution already control for this confounder.

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

Table 1 implements these adjustments to Barceló and Labzina (2020) fixed-effects model[3] by explicitly showing how the estimated coefficient, t-statistics, and confidence intervals at the 95 per cent confidence level would look under varying levels of postulated confounding. We show this relative to the confounder of Twitter suspension activity. Taking a conservative scenario where a plausible confounding factor would be *three* times as strong as the observed number of suspended Twitter accounts, the estimated effect of terrorist attacks in Europe or the United States on the number of followers of IS-related Twitter accounts is −0.069, with an associated t-statistic of 7.90. When attacks take place in other countries, the estimated effect is −0.026, with an associated t-statistic of −12.28. Both estimates remain statistically significant at the 99 per cent confidence level under this level of confounding. Table 1 also reveals that a confounder would need to be *six* times as strong as that of the observed number of suspended Twitter accounts to reduce the estimates for the effect of the number of victims in Europe or the United States. At the same time, Table 1 shows that for attacks outside Europe and the United States, a confounder would need to be *eleven* times as strong to reduce the estimates below conventional levels of statistical significance in the same manner.

The Direct Effect of Terrorist Attacks on Follower Loss for IS-Related Accounts

Rather than considering it as a confounder, we will now treat Twitter suspension efforts as a plausible intermediate factor between terrorist attacks and the subsequent follower loss of IS-related Twitter accounts. We therefore reevaluate our findings using the g-estimator—a two-stage regression estimator—that allows scholars to assess posttreatment mediators (Acharya, Blackwell, and Sen 2016). It should be noted that the ACDE is identified even in the face of intermediate confounders. This methodological approach allows us to account for postattack factors, such as Anonymous and Twitter joint suspension efforts, without inducing bias in our model estimates (Montgomery, Nyhan, and Torres 2018).

In the first stage, the method estimates a model with both pretreatment—the intensity of terrorist attacks—and posttreatment covariates—Anonymous reporting activities and Twitter suspension efforts. Then, it reestimates the dependent variable—the log of the number of followers—by removing from it the effects of the mediating variables of interest—Anonymous reporting activities and Twitter suspension efforts—which are set to zero mediating effect. Lastly, it calculates the estimated effect of the treatment—the intensity of terrorist attacks—on this *de-mediated* outcome. In other words, by accounting for Anonymous and Twitter suspension efforts with the sequential g-estimator, we can recover the controlled direct effect of terrorist attacks on follower loss regardless of the social media efforts to fight IS-linked accounts.

Table 2 reports the g-estimator of the number of deaths in terrorist attacks on the log of the number of followers of IS-related Twitter accounts using the sequential g-estimator. The estimated coefficient for victims in Europe or the US is −0.11, which is reliable and substantively meaningful. Similarly, the estimated coefficient for victims outside Europe or the US is −0.03, which is also reliable and substantively meaningful. Both estimates indicate that an increase in the number of victims from IS-related terrorist attacks leads to a subsequent loss in the number of followers of IS-related Twitter accounts. As a reminder, these models use the fixed-effect estimator, implying that the estimator relies on within- account variations of followers over time. As such, it compares observations within the same account at times where the level of deaths from attacks differs. In this way, we are able to only use information from accounts that are observed more than once.

According to the model, the direct effect of a terrorist attack that produces 50 deaths in Europe or the United States is associated with a decline from 5.2 per cent to 6.95 per cent in the number of followers in IS-related Twitter accounts, taking into account the effect that is not mediated by

[3]See Model 1.4 of Table 2 in Barceló and Labzina (2020, 1553).

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

Table 1. Regression results with OVB relative to the observed number of suspended accounts by Twitter

Size of the bias relative to the variable: *N* of suspended accounts by Twitter	$R^2_{D\sim Z\mid \mathbf{X}}$	$R^2_{Y\sim Z\mid \mathbf{X},D}$	Treatment	Estimate (SE)	Value	Lower CI (95%)	Upper CI (95%)	Significance level
1 × *N* of suspended accounts	0.037	0.0004	Victims in the US/Europe	−0.100 (0.008)	−11.85	−0.117	−0.083	***
1 × *N* of suspended accounts	0.018	0.0003	Victims outside the US/Europe	−0.031 (0.002)	−14.91	−0.035	−0.027	***
2 × *N* of suspended accounts	0.074	0.0007	Victims in the US/Europe	−0.085 (0.009)	−9.89	−0.102	−0.068	***
2 × *N* of suspended accounts	0.036	0.0007	Victims outside the US/Europe	−0.029 (0.002)	−13.60	−0.033	−0.025	***
3 × *N* of suspended accounts	0.111	0.0011	Victims in the US/Europe	−0.069 (0.009)	−7.90	−0.086	−0.052	***
3 × *N* of suspended accounts	0.055	0.0010	Victims outside the US/Europe	−0.026 (0.002)	−12.28	−0.030	−0.022	***
4 × *N* of suspended accounts	0.148	0.0014	Victims in the US/Europe	−0.053 (0.009)	−5.92	−0.070	−0.035	***
4 × *N* of suspended accounts	0.072	0.0014	Victims outside the US/Europe	−0.024 (0.002)	−10.96	−0.028	−0.019	***
5 × *N* of suspended accounts	0.185	0.0018	Victims in the US/Europe	−0.036 (0.009)	−3.92	−0.054	−0.018	***
5 × *N* of suspended accounts	0.091	0.0017	Victims outside the US/Europe	−0.021 (0.002)	−9.63	−0.025	−0.017	***
6 × *N* of suspended accounts	0.222	0.0022	Victims in the US/Europe	−0.018 (0.010)	−1.90	−0.036	0.001	*
6 × *N* of suspended accounts	0.109	0.0021	Victims outside the US/Europe	−0.018 (0.002)	−8.31	−0.023	−0.014	***
7 × *N* of suspended accounts	0.259	0.0026	Victims in the US/Europe	0.0013 (0.010)	0.133	−0.018	0.020	n.s.
7 × *N* of suspended accounts	0.127	0.0024	Victims outside the US/Europe	−0.015 (0.002)	−6.98	−0.020	−0.011	***
8 × *N* of suspended accounts	0.296	0.0029	Victims in the US/Europe	0.022 (0.010)	2.18	0.002	0.041	n.s.
8 × *N* of suspended accounts	0.145	0.0028	Victims outside the US/Europe	−0.013 (0.002)	−5.64	−0.017	−0.008	***
9 × *N* of suspended accounts	0.0033	0.0033	Victims in the US/Europe	0.043 (0.010)	4.25	0.023	0.063	n.s.
9 × *N* of suspended accounts	0.163	0.0031	Victims outside the US/Europe	−0.010 (0.002)	−4.30	−0.014	−0.005	***
10 × *N* of suspended accounts	0.370	0.0037	Victims in the US/Europe	0.066 (0.010)	6.34	0.046	0.087	n.s.
10 × *N* of suspended accounts	0.182	0.0035	Victims outside the US/Europe	−0.007 (0.002)	−2.96	−0.011	−0.002	***
11 × *N* of suspended accounts	0.407	0.0041	Victims in the US/Europe	0.091 (0.011)	8.47	0.070	0.112	n.s.
11 × *N* of suspended accounts	0.200	0.0038	Victims outside the US/Europe	−0.004 (0.002)	−1.62	−0.008	0.001	n.s.

Notes: CI = confidence interval; n.s. = not significant.

Table 2. ACDE of terrorist attacks on the number of followers of IS-related Twitter accounts

	DV: number of followers (log scale)		
	Discount rate: 50%	Discount rate: 75%	Discount rate: 100%
Explanatory variables:			
Number of victims (US and Europe) ('00)	−0.11*** (0.011)	−0.13*** (0.013)	−0.15*** (0.015)
Number of victims (outside US and Europe) ('00)	−0.03*** (0.002)	−0.03*** (0.002)	−0.04*** (0.002)
Mediators:			
N of reported accounts by Anonymous ('0000)	Yes	Yes	Yes
N of suspended accounts by Twitter ('000)	Yes	Yes	Yes
FE user account	Yes	Yes	Yes
FE time	Yes	Yes	Yes
N total	217,844	217,844	217,844
N of accounts	8,300	8,300	8,300
Time range (in days)	127	127	127

Notes: Explanatory variables and controlled intermediate variables are de-meaned within user accounts. This allows us to combine the direct controlled effect method with the user-account fixed-effect estimator. Controlled intermediate variables are de-mediated, that is, they are set to 0. Discount rates refer to the discounting factors applied to the daily number of deaths, which is then used to create the cumulative number of deaths (see Barceló and Labzina 2020, 1546). DV = dependent variable. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Twitter suspension efforts. For the average account, which has 348 followers, a large-scale terrorist attack of 50 deaths would imply a decrease resulting in between 324 and 330 followers after the attack. We can also calculate that such an attack would reduce the total number of followers of IS-related accounts in Twitter by a sizable amount of 149,400–199,200 followers.[4] Similarly, an attack outside of Europe and the United States that produces 100 deaths[5] would lead to a follower loss of between 2.95 per cent and 3.92 per cent. For the average account, this reduction would shift the number of followers from 348 to either 338 or 335 depending on the discount rate. Aggregating this loss of followers across accounts, this is equivalent to 83,000 or 107,900 fewer followers of IS-related accounts across the entirety of Twitter.[6]

The causal interpretation of the g-estimator requires an assumption of sequential unconfoundedness, which represents two simultaneous *no omitted variables* assumptions. First, for the identification of the g-estimator, we require "no omitted variables for the effect of treatment on the outcome, conditional on the pretreatment confounders" (Acharya, Blackwell, and Sen 2016, 519). This assumption is plausible in the context of our research design. Within a narrow time window of a 127-day period, the exact timing of a terrorist attack is the joint result of unrelated processes, including but not limited to counterterrorist strategies, local festivities and celebrations, and the organizational capabilities of the authors of the attack. We believe that it is unlikely that a single variable causally affects both the timing of attacks and Twitter followers of IS accounts once we account for overall time trends in the model within this narrow time window.

The second *no omitted variable* assumption requires "no omitted variables for the effect of the mediator on the outcome, conditional on the treatment, pretreatment confounders, and intermediate confounders" (Acharya, Blackwell, and Sen 2016, 519). In our context, the g-estimator would identify an ACDE if there is no omitted variable causally connected to both changes in Twitter suspension activity and the follower shift in IS-related Twitter accounts before and after the attacks. While we believe that our fixed-effects g-estimator model is quite stringent with regards to identification, we cannot assure that we are covering all the intermediate confounding variation. As indicated earlier, we believe that it is more productive to empirically

[4]These values refer to followers but not unique users. We also use the total number of accounts observed in the fixed-effects models for calculating this value.

[5]We use 100 deaths, rather than 50, in these simulations because attacks with more than 100 deaths are more common outside of Europe and the United States.

[6]As earlier, these values refer to followers but not unique users.

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

investigate the fragility of the g-estimator as compared to deviations of the zero sequential unconfounded bias than engaging in a speculative debate on that amount.

We evaluate the sensitivity of the g-estimator to violations of the sequential unconfoundedness assumption. Figure 1 shows the estimated ACDE for terrorist attacks on European and North American soil at varying magnitudes of violation of the assumptions behind the g-estimator.[7] To provide a meaningful mapping of the sensitivity of the ACDE, the range of the reported violations on the x-axis are benchmarked to the observed correlation between the mediator and the outcome variable. Figure 1 reveals that it would require a correlation between the mediator and outcome errors of −0.16 for the estimated effect of attacks in Europe or the United States to lose statistical significance at the 95 per cent confidence level. This is equivalent to a correlation 26 times as *strong* as the observed correlation between the mediator and the outcome variable. Furthermore, Figure 1 also shows that the estimated ACDE for attacks outside Western countries is robust to virtually any meaningful correlation between the mediator and outcome errors. Thus, we conclude that it is unlikely there exists a confounder as large as could alter our substantive interpretation of causal direct effects.

Replatforming?

Barceló and Labzina's (2020) main theoretical argument is that terrorist attacks demobilize terrorist supporters, which is reflected in the decline of followers of IS accounts. Hansen argues that leaving Twitter might be the result not of demobilization, but of a move from Twitter to a replaceable and analogue platform: Telegram Messenger. Again, Hansen provides no empirical evidence for this claim.

In this section, we contend that the replatforming of users from Twitter to Telegram is unlikely to fully absorb the observed reduction of Twitter followers for two reasons: first, the distinct nature of Twitter and Telegram makes them complementary, not exchangeable, platforms; and, second, empirical patterns are inconsistent with a process of replatforming. We elaborate each in turn.

In its current form, Twitter is not replaceable by Telegram. In the early days of Telegram in 2015, IS began using this platform to broadcast propaganda messages to an unlimited number of users—analogously to Twitter propaganda channels. However, Telegram blocked nearly 80 IS-related accounts after the Paris attacks in November 2015 (Novet 2015). While Telegram kept blocking IS-related public broadcast channels, it also announced that it could not intercept private communication via chat rooms (Berger and Perez 2016), and chat rooms became IS's main Telegram tool. Yet, unlike public broadcast channels, chat rooms are not searchable, users must "join" a chat group to view its content and provide a specific web address to become a group member (Berger and Perez 2016).

Due to these limitations, Telegram has a different purpose and audience than Twitter. While Twitter is used for public broadcasting and interaction with the general public and noncore supporters, as well as recruiting new members, Telegram is used to discuss sensitive matters, such as the organization and coordination of terrorist attacks or travel to IS-held territory, among IS strong believers. According to a revealing IS-related Telegram post cited in Berger and Perez (2016, 19):

> Telegram is not a media platform for dawa [proselytizing] to all Muslims and the West. No one will enter your channel except the Ansar [IS supporters] who already know the truth. Or your enemies to report you. Rarely would you find someone from general public following you. That's why our main platform is [w]here the General Public is found. Like on Twitter and Facebook. We are here for Dawa. Not to entertain each other and talk to each other…. Let Telegram be like an archive.

[7]In the sensitivity models reported here, we reestimate those with a discount rate of 100 per cent, though using 50 per cent and 75 per cent discount rates does not change our conclusions (see the Online Data Replication files).

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

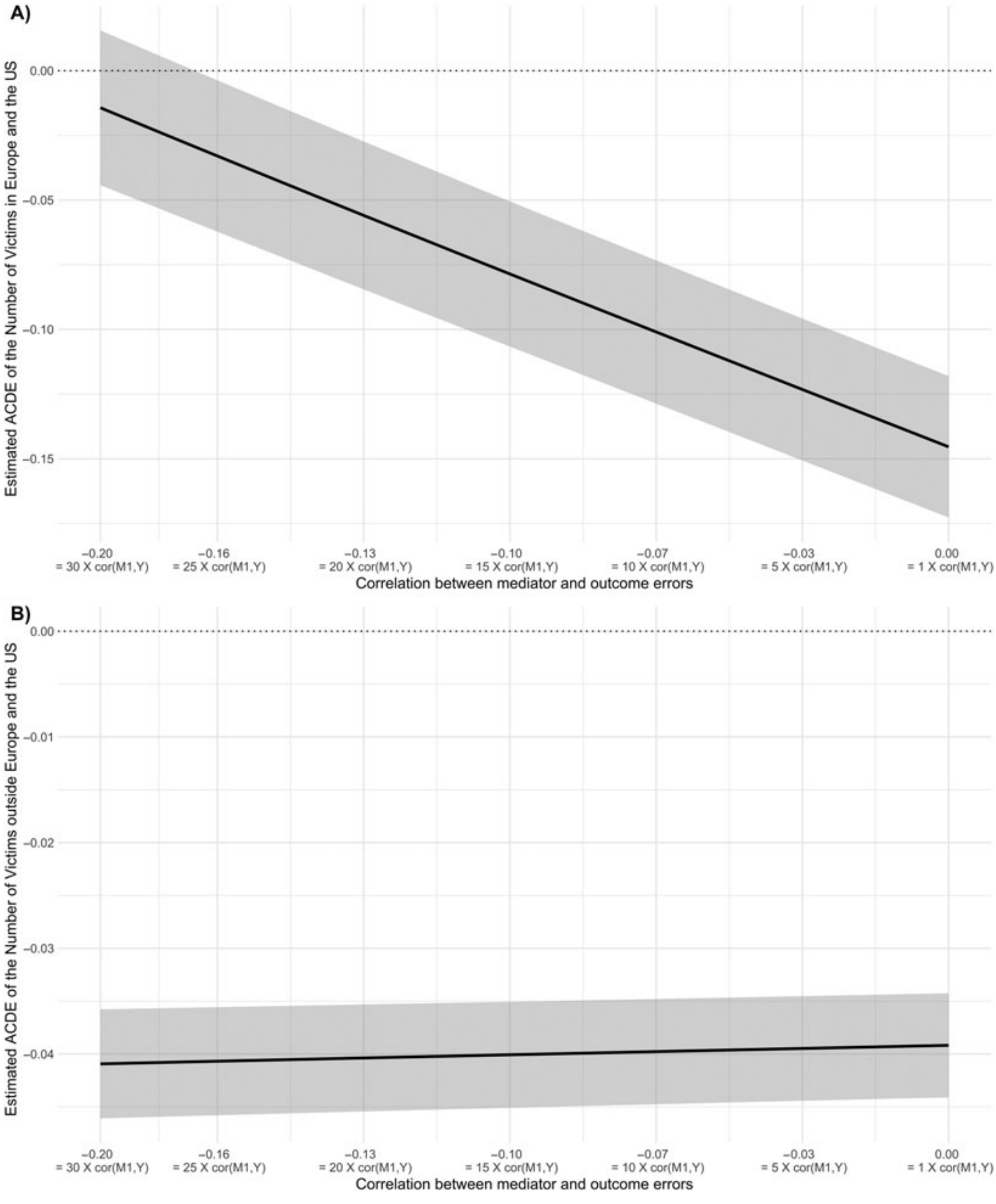

Fig. 1. Sensitivity analysis of the ACDE of terrorist attacks on the number of followers in IS-related Twitter accounts.

This cross-platform synergy means that the number of IS-related accounts and followers on Twitter is likely to be positively correlated with the number of IS-related accounts on Telegram. A reduction in Twitter followers is unlikely to simply result from users who move from Twitter to Telegram because they serve different functions. It is more likely that the individuals' presence in one is positively—not negatively—correlated with their presence in the other.

Secondly, we also contend that empirical patterns are inconsistent with replatforming. Having ruled out the Twitter suspension and Anonymous reporting efforts in the previous sections, there

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

are two reasons that could plausibly motivate users' replatforming after attacks: a change in beliefs about IS; and/or heightened perceptions of prosecution among IS supporters.

First, the change in beliefs is unlikely to lead to a migration from Twitter to Telegram exclusively. As shown earlier, IS-related Telegram pushed users to create and maintain accounts in both platforms, that is, Twitter and Telegram, simultaneously. Therefore, those who might change their behavior because of an increased belief in IS's ideas would, if anything, change their behavior to follow IS's recommendations, not contradict them. This is the first reason change in pro-IS beliefs is unlikely to lead to replatforming.

Secondly, users could also move from Twitter to Telegram after an attack if they are of a belief that Telegram is under less surveillance than Twitter. This argument is analogous to a deterrence effect by which users might decide to unfollow IS-related accounts on Twitter by keeping their engagement in the "dark" (that is, Telegram). Following our original empirical strategy, we suggested that "those followers located in the countries with a more credible threat to prosecute or retaliate against online activists, or the countries with greater national material capabilities, should be more likely to drop from following IS-related accounts in the aftermath of an attack" (Barceló and Labzina 2020, 1544). We found no evidence suggesting that the dropout is concentrated in countries with strong counterterrorist capabilities, indicating that the postattack reduction in followers is unlikely to arise from safety concerns.

Discussion

Every additional terrorist attack reveals information about the character and the goals of the organization perpetrating the attack. Some individuals who once supported or sympathized with an organization may come to believe that the violence has gone too far and that they cannot cope with new levels of brutality, leading to a process of disengagement. Empirically, some qualitative studies evidence this effect in the context of far-right movements (Bjørgo 2011) and domestic terrorist organizations (Alonso 2011; Reinares 2011). Our article provides the first quantitative evidence for the effect of terrorist attacks in the case of IS—one of the bloodiest contemporary terrorist organizations. We also joined an emerging trend in social science of using social media data to shed light on fundamental questions (for example, Barberá et al. 2015), including conflict research (Zeitzoff 2011).

Hansen's critique describes some plausible perils about using social media data to evaluate the consequences of terrorist attacks. We agree that these limitations need to be considered when designing research projects using social media data. However, we disagree with Hansen's view that social media data cannot be used to advance knowledge in conflict and terrorism research. Our original article, Hansen's comment, and this response in turn provide a roadmap for identifying potential limitations and some guidance on how to deal with challenges that arise with social media data. We focus our discussion on two major concerns: the changing composition of the sample; and the (non)replaceable nature of media platforms.

Changing composition of the sample

The methodological literature on social media data has paid specific attention to the issue of representativeness. Facebook and Twitter users are, for instance, younger, better educated, more liberal, more interested in politics, more likely to hold extreme political views, and more likely to live in urban areas than nonusers (Barberá and Rivero 2015; Mellon and Prosser 2017). This might raise concerns for the external validity of studies that seek to make inferences about the general population using data from these platforms.

Less attention has been paid to how changes in the composition of social media samples over time may bias inferences. The composition of a sample from social media may shift over time in nonrandom ways. Scholars often use Facebook or Twitter data to infer how key attributes, such as

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

network density, the content of Twitter posts, and the number of followers, change over time as a response to naturally occurring events, such as terrorist attacks (Barceló and Labzina 2020; Fischer-Preßler, Schwemmer, and Fischbach 2019; Garcia and Rimé 2019), electoral results (Mitts 2019), foreign policy (Jones and Silver 2020), and even pandemics (Huang et al. 2020). However, social media samples are dynamic, in the sense that users may deactivate and activate accounts, or may intensify their activity, as a function of salient events, including the event of interest. These changes in the composition of the sample add an inferential challenge because treatment effects integrate two mechanisms: within-individual behavioral changes; and shifts in the composition of the sample.

As researchers seek to remove changes in the composition of the samples over time from their models, our exchange with Hansen highlights the importance of two specific choices of model specification that would allow scholars to appropriately identify a causal effect in similar circumstances: (1) user-accounts fixed effects; and (2) narrow time periods. As we elaborated earlier, user-account fixed effects with pretreatment outcome measures ensure a stable sample before and after the salient event of interest. Neither the resurgence of suspended accounts nor the suspension of accounts based on account-level observables is a concern in models employing user-account fixed effects. Second, analyses over a narrow time period—that is, days or weeks around the event, rather than years—ensure that the sample of users does not differ extensively during the time period and also increase the likelihood that pretreatment outcome measures better resemble their counterfactual—posttreatment outcome measures in the absence of the treatment. In studies where the outcome of interest is social media data gathered from extremist organizations, changes in the composition of the sample may be particularly concerning because of the active prosecution of these users by some social media platforms like Twitter. When possible, we strongly encourage scholars to collect information about platform deactivation activities to evaluate their impact on the estimated effects using existing analytical approaches.

Specifically for studies on conflict and terrorism using Twitter data, we suggest scholars collect information about Twitter suspension efforts and Anonymous reporting activities when establishing their research designs. While we think that these variables are posttreatment, we have shown how these measures can be useful to de-mediate the estimated effects from variance that comes from suspensions. To do this, we suggest researchers report their estimates together with the ACDE of terrorist attacks to show the estimated effects on outcomes measured using social media while not going through the pathway of suspensions. We understand that the sequential unconfoundedness assumption raises an additional practical challenge. In settings where an unmeasured factor could confound the association between terrorist attacks and the outcome, or Twitter suspension efforts and the outcome, scholars should report their g-estimators together with sensitivity analyses using existing analytical approaches (Acharya, Blackwell, and Sen 2016). With sensitivity analyses, scholars can present imperfect results while transparently revealing how susceptible the estimates are to violations of sequential unconfoundedness. We envision that our efforts in the study of terrorism using Twitter data could be generalized to other studies where social media platforms take an active role in deactivating accounts associated with large-scale events.

The replaceable versus nonreplaceable nature of social media platforms

Social media data are not necessarily replaceable across platforms. While some organizations and users may consider Twitter, Facebook, Telegram, WeChat, WhatsApp, and the myriad other available social media platforms as fully interchangeable, others may consider them to be single, unique forms of communication that serve different purposes and have different target audiences. The simplistic argument that users can replace one platform with another with no further consequences is a nonobvious claim, which might end up being inaccurate. For instance, as discussed earlier, IS encouraged their core users to use multiple platforms simultaneously to achieve

https://doi.org/10.1017/S0007123421000582 Published online by Cambridge University Press

different goals at the same time, that is, while Twitter was used to interact with the general audience and recruit new members, Telegram was more often used to discuss sensitive, specific organizational and coordination matters. However, we acknowledge that some social media platforms are capable of replacing one another. For example, one could think that organizations might tend to use either WhatsApp or WeChat, yet rarely both at the same time, as they are almost identical functionally.

We encourage scholars who employ social media data in their research to evaluate social media platforms through this lens of their relations with other platforms. These relationships might have critical implications for the interpretation of empirical findings in two distinct directions.

First, if researchers employ social media data from a single platform that is connected in a *substitute* relationship with another unobserved platform, then behavioral changes in the observed platform might simply indicate a behavioral transfer to the unobserved platform. Hence, users' time could be reallocated from one platform to another, and assuming activity remains constant in both platforms, users might transfer some posts that would have been posted in the old platform to the new, and vice versa.

Secondly, at the same time, if researchers employ social media data from a single platform that is connected in a *complementary* relationship to another platform, then deplatforming is less of a concern because users are unlikely to transfer their behaviors across platforms. Instead, scholars might focus their discussion on evaluating whether their findings from one platform could be applied to data from other platforms and what implications this may have for the interpretation of their findings. For instance, if an organization uses Facebook to broadcast information and Twitter to recruit new supporters, empirical findings might not travel across platforms because the data mean different things on different platforms.

Before conducting a study such as ours, we encourage scholars to understand how each platform of study is embedded in a network of platforms. Scholars should then discuss the chances of deplatforming, replatforming, or simultaneous use and the possible implications of their data-generation process, as well as the implications of their findings. Furthermore, we strongly believe that, while challenging from a logistical perspective, cross-platform studies that pool together information across several platforms in the same study will help us highlight these patterns of interaction across platforms and constitute a promising avenue for future research.

In summary, our exchange with Hansen has allowed us to better diagnose some of the perils of using social media data that are likely to be relevant across numerous investigations across the social sciences. At the same time, it has enabled us to further strengthen our original work by providing additional robustness checks and a more thorough evaluation of our empirical findings. We hope that this article not only serves as a response to Hansen's speculations, but also provides a roadmap for future scholars to address these concerns. If approached through this lens of skepticism and rigorous sensitivity testing, scholars can keep relying on social media data in the study of terrorism and related fields.

Data Availability Statement. Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/B02HVQ

Acknowledgments. We are grateful to the editor, Rob Johns, and an anonymous reviewer for helpful comments and suggestions.

Author Contributions. Joan Barcelo contributed to: conceptualization; formal analysis; methodology; writing—original draft; and writing—review and editing. Elena Labzina contributed to: data curation; and writing—review and editing). Authors listed in order of contribution.

Financial Support. None.

Competing Interests. None.

References

Acharya A, Blackwell M, and Sen M (2016) Explaining causal findings without bias: detecting and assessing direct effects. *American Political Science Review* **110**(3), 512.

Alonso R (2011) Why do terrorists stop? Analyzing why ETA members abandon or continue with terrorism. *Studies in Conflict & Terrorism* **34**(9), 696–716.

Barberá P and Rivero G (2015) Understanding the political representativeness of Twitter users. *Social Science Computer Review* **33**(6), 712–729.

Barberá P et al. (2015) Tweeting from left to right: is online political communication more than an echo chamber? *Psychological Science* **26**(10), 1531–1542.

Barceló J and Labzina E (2020) Do Islamic State's deadly attacks disengage, deter, or mobilize supporters? *British Journal of Political Science* **50**(4), 1539–1559.

Barceló J and Labzina E (2021) Replication Data for: "Islamic State's Terrorist Attacks Disengage Their Supporters: Robust Evidence from Twitter," https://doi.org/10.7910/DVN/B02HVQ, Harvard Dataverse, V1.

Berger JM and Perez H (2016) The Islamic State's Diminishing Returns on Twitter: How Suspensions Are Limiting the Social Networks of English-Speaking ISIS Supporters. George Washington University. Available from https://bit.ly/3AYxO6K

Bjørgo T (2011) Dreams and disillusionment: engagement in and disengagement from militant extremist groups. *Crime, Law and Social Change* **55**(4), 277–285.

Cinelli C and Hazlett C (2020) Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(1), 39–67.

Cinelli C, Ferwerda J, and Hazlett C (2020) Sensemakr: Sensitivity Analysis Tools for OLS in R and Stata. Working paper. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588978

Fischer-Preßler D, Schwemmer C, and Fischbach K (2019) Collective sense-making in times of crisis: connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior* **100**, 138–151.

Garcia D and Rimé B (2019) Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science* **30**(4), 617–628.

Hansen T (2022) The perils of estimating disengagement effects of deadly terrorist attacks utilizing social media data. *British Journal of Political Science*. doi: 10.1017/S000712342100017X.

Huang X et al. (2020) Twitter reveals human mobility dynamics during the COVID-19 pandemic. *PloS One* **15**(11), e0241957.

Jones NM and Silver RC (2020) This is not a drill: anxiety on Twitter following the 2018 Hawaii false missile alert. *American Psychologist* **75**(5), 683.

Mellon J and Prosser C (2017) Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. *Research & Politics* **4**(3), 2053168017720008.

Mitts T (2019) From isolation to radicalization: anti-Muslim hostility and support for ISIS in the West. *American Political Science Review* **113**(1), 173–194.

Montgomery JM, Nyhan B, and Torres M (2018) How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science* **62**(3), 760–775.

Novet J (2015) One app maker has shut down almost 80 secret channels used by ISIS to communicate. *Business Insider,* November 18. https://www.businessinsider.com/telegram-cracks-down-on-isis-related-channels-2015-11

Reinares F (2011) Exit from terrorism: a qualitative empirical study on disengagement and deradicalization among members of ETA. *Terrorism and Political Violence* **23**(5), 780–803.

Zeitzoff T (2011) Using social media to measure conflict dynamics: an application to the 2008–2009 Gaza conflict. *Journal of Conflict Resolution* **55**(6), 938–969.

Cite this article: Barceló J, Labzina E (2022). Islamic State's Terrorist Attacks Disengage Their Supporters: Robust Evidence from Twitter. *British Journal of Political Science* **52**, 1490–1501. https://doi.org/10.1017/S0007123421000582