Online Appendix for "Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes"*

Ala' Alrababa'h[†] William Marble[‡] Salma Mousa[§] Alexandra Siegel[¶]

American Political Science Review
April 2021

*An expanded appendix, which includes additional survey experiment analyses, is available on the *APSR* Dataverse at `https://doi.org/10.7910/DVN/2JKWNS`.

[†]Department of Political Science, Stanford University, and Immigration Policy Lab, Stanford University and ETH Zurich. Email: alaa@stanford.edu. ORC ID: https://orcid.org/0000-0001-5762-8892

[‡]Department of Political Science, Stanford University. Email: wpmarble@stanford.edu. ORC ID: https://orcid.org/0000-0001-9352-5540

[§]Department of Political Science, Stanford University, and Immigration Policy Lab, Stanford University and ETH Zurich. Email: smousa@stanford.edu. ORC ID: https://orcid.org/0000-0002-1482-4276

[¶]Department of Political Science, University of Colorado Boulder, and Immigration Policy Lab, Stanford University and ETH Zurich. Email: alexandra.siegel@colorado.edu. ORC ID: https://orcid.org/0000-0003-0792-7813

Contents

A Hate Crimes Analysis

A.1 Data Collection

To gather data on hate crimes, we submitted Freedom of Information requests (FOI) to every police department in England in April 2018. We requested a dataset consisting of every hate crime that was reported to the department between January 2015 and April 2018, along with information including the date, location, motivation for the crime, and demographic information about the victim. We include a police jurisdiction in our analysis if its response provided sufficient information for us to calculate the total number of hate crimes reported in the jurisdiction for each month. We obtained usable data from from 25 police jurisdictions out of the 39 contacted, and 936 month-police force observations. Hate crimes themselves cover a range of offenses. Common violations include harassment, aggravated common assault, criminal damage to vehicles, and aggravated public fear, alarm, or distress. In order to be classified as a hate crime, police should have a clear indication that the perpetrator targeted the victim mainly on the basis of their religious, racial, sexual, or abilities-based identity.

In our analysis, we use all reported hate crimes. We requested data on hate crimes broken down by victim religion and ethnicity, but the responses were inconsistent. In some cases, police departments do not collect this information; in others, they began collecting it near the end of the study period. As a result, we include all reported hate crimes. The focus on all hate crimes should still reflect trends driven by anti-Muslim incidents: the Home Office reports that 76% of hate crimes perpretrated from January 2017 to January 2018 were religiously or racially motivated.[1] Of these crimes, 52% were categorized as anti-Muslim in particular (BBC News, 2018).

Our main outcome variable is an annualized hate crime rate per thousand residents. For instance, a police jurisdiction with a population of 100,000 that experiences 10 hate crimes in a given month has an annual hate crime rate of (10 / 100,000) $\times$ 1,000 $\times$ 12 = 1.2 hate crimes per thousand residents in that month.[2] The dependent variable ranges from 0 to a maximum of 4.342, with a mean of 0.951 and standard deviation of 0.767.

A.2 Treatment Assignment

We consider the Merseyside police force — which covers Liverpool — to be treated after Salah's official signing in June 2017. Merseyside is a metropolitan county that encompasses both Everton F.C. and Liverpool F.C. fans.[3] While a Salah effect is likely to be most pronounced after his stellar performances with the team in late 2017, we choose his signing date as the start of treatment for two reasons. First, any other cutoff would be somewhat arbitrary, whereas there is a clear justification for choosing June 2017. Second, when Salah was signed, his transfer fee constituted a club record, stoking interest in the player among the club's fans. Figure A-1 shows that public interest in Salah — as measured by Google searches in the U.K. — spiked shortly after he was signed in the summer of 2017 and then began to steadily increase afterwards through mid-2018. In Appendix C we show that mentions of Salah in *Liverpool Echo* headlines follow a similar trend. We also discuss other events relevant to Islamophobia that occurred around this time — which could complicate interpretation

[1] Note that the same offense can be categorized as both racially and religiously motivated.

[2] Any other normalization procedure would yield identical results, up to a multiplicative constant.

[3] A backlash among Everton fans would dilute any treatment effects for the hate crime analysis, biasing against finding an effect.

of our estimates — in the section "Robustness and Generalizability Tests" of the main text and in Appendix A.5.

Figure A-2 plots the raw time series data for each police force, with Merseyside highlighted. In the pre-treatment period, hate crimes are relatively common in Merseyside. Averaging over all pre-treatment observations, the hate crime rate in Merseyside is higher than 19 of the other 24 police forces in the data.

A.3 Research Design

Our goal is to estimate a counterfactual quantity: the predicted trajectory of hate crimes in Merseyside had Salah not joined Liverpool. A number of methods have been developed for this task, including two-way fixed effects models, interactive fixed effects, the synthetic control method, and matrix completion methods (Abadie, Diamond and Hainmueller, 2010; Doudchenko and Imbens, 2016; Xu, 2017; Athey et al., 2018). Roughly speaking, these methods attempt to impute the unobserved outcomes in the post-treatment period by first looking for structure in the pre-treatment data that generates good predictions of the treated unit's outcomes in the pre-treatment period. The same structure is then applied to the post-treatment periods to generate estimates of the counterfactual potential outcomes for the treated unit. To obtain an estimate of the treatment effect on the treated unit, we simply take the difference between the observed outcome for the treated unit in the post-treatment period and the imputed counterfactual outcome. Therefore, if there are T post-treatment periods, we obtain T treatment effect estimates. In addition, we compute the treatment effect averaged over the T post-treatment periods as a simple summary of the treatment effect.

This method outperforms others in approximating the outcome in Merseyside prior to the treatment period, so arguably generates a more suitable counterfactual estimate than others. This method attempts to find a low-dimensional matrix structure in the data by minimizing the mean squared error between the observed outcomes and the outcomes predicted by another (low-rank) matrix. To avoid overfitting, the procedure penalizes the complexity of the matrix by adding a penalty term proportional to the nuclear norm of the matrix, with the scaling factor chosen via leave-one-out cross-validation.[4]

Statistical inference in the setting of a single treated unit is challenging. Standard methods for computing standard errors based on asymptotic theory obviously do not apply. We implement three complementary approaches to inference: the nonparametric bootstrap, a permutation-based method, and a placebo analysis leveraging other types of crimes.

First, by repeatedly resampling control units and re-estimating the model, we generate a bootstrap distribution and standard error for the treatment effect estimator. We can then compute a standard error by taking the standard deviation of bootstrap estimates and obtain confidence intervals by taking the appropriate quantiles of the bootstrap distribution.

Second, we reshuffle units' treatment status to generate a reference distribution for the estimator. For each control unit, we pretend it was in fact treated and estimate the "treatment effect" on the placebo treated unit. By construction, there is 0 treatment effect for these units (since they were not actually treated), so this procedure generates a distribution of the treatment effect estimator under the sharp null

[4] In each cross-validation iteration, we omit one pre-treatment observation for the treated unit. We then select the penalization parameter that produces the smallest mean-squared prediction error for the held out observations. Because we have 30 pre-treatment periods for Merseyside, we choose $k = 30$-fold cross validation in the `gsynth` software.

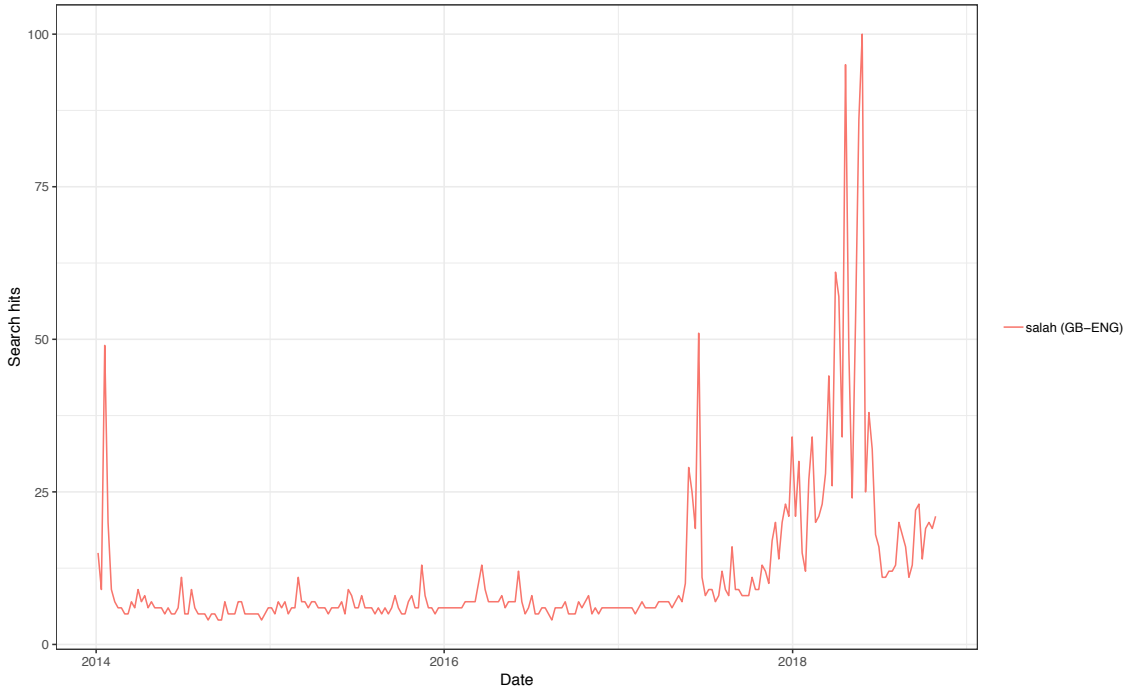


Figure A-1: Normalized Google Searches for "Salah" in the UK (2014-2018)

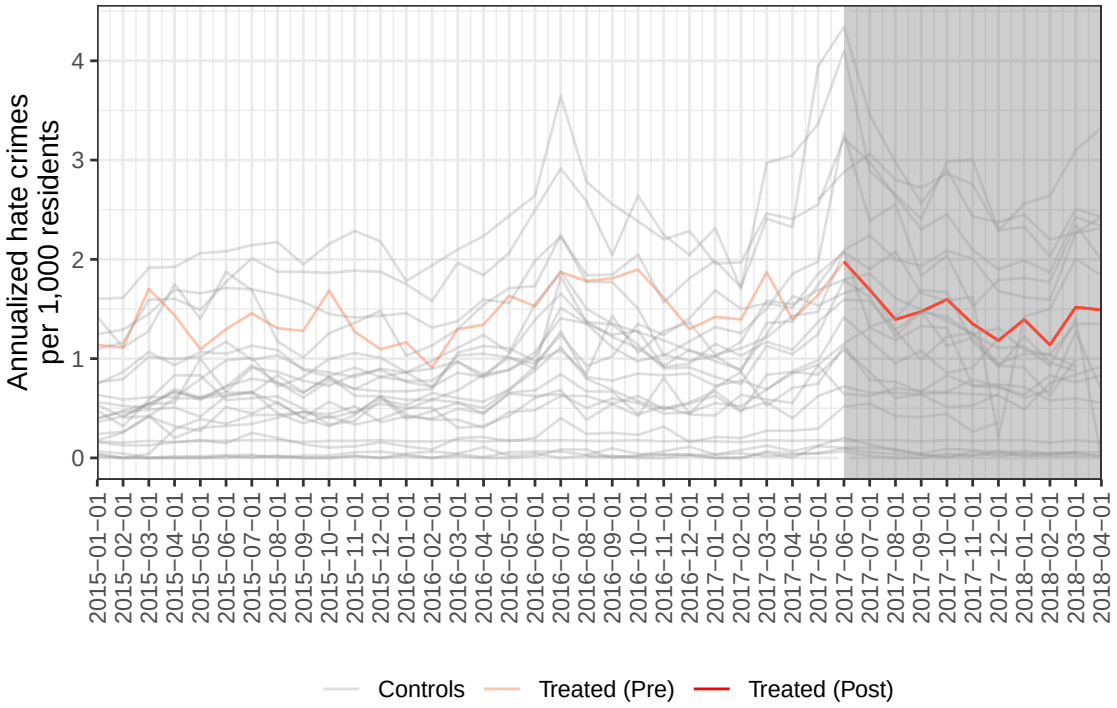


Figure A-2: Hate crime data across police jurisdictions. The red line shows hate crimes reported by the Merseyside police force.

of no treatment effect in any period, for any unit. We can then take the actual Merseyside estimates and compare them to the null distribution to generate a permutation-based p-value. This method of inference is proposed by Abadie, Diamond and Hainmueller (2010).[5]

Third, we conduct a placebo test using other types of crime that are unlikely to be affected by changes in anti-Muslim sentiment. We collected data from the U.K. Home Office on crime at the police jurisdiction level.[6] These data are formatted in a standard set of 14 crime types (which does not include hate crimes), such as shoplifting, robbery, possession of weapons, drugs, and so on. There is little reason to believe that these crimes would be affected by a decrease in anti-Muslim sentiment. If we find a significant decrease in these crimes in Merseyside after Salah was signed, it would indicate that a decrease in hate crimes may be explained as part of a more general trend. To conduct this test, we re-run the matrix completion analysis on each of the placebo outcomes, then compare the estimated effect sizes (normalized by the pre-treatment mean for each outcome).

A.4 Generalized Difference-in-Difference

An alternative method of measuring the effect of Salah on hate crimes is to employ a generalized difference-in-differences framework by estimating a two-way fixed-effects (TWFE) regression of the form

$$Y_{it} = \tau D_{it} + \delta_i + \gamma_t + \epsilon_{it}, \tag{A-1}$$

where D_{it} is an indicator that switches on for Merseyside in the post-treatment period, δ_i and γ_t are unit and month fixed effects, respectively, and ϵ_{it} is a mean-zero error term. In this framework, given parallel trends for the treated and untreated units in the absence of treatment, τ is the the ATT.

Table A-1 presents the main regression results. The first column reports the plain TWFE model, the second adds police force-specific linear time trends, and the third and fourth add population weights. All specifications give similar results, showing that there was a decrease in the hate crime rate in Merseyside after Salah was signed. The estimates are in the range of -0.2, which is very similar to the estimated ATT yielded by the matrix completion method, which was -0.275.

In all the regression models, the estimates appear to be significant. However, with only a single treated unit, the standard errors may not be reliable. We therefore undertake an alternative form of inference, whereby we randomly assign a single unit to be treated, with treatment beginning in a randomly chosen month that is at least 4 months after the first observations in our dataset and as late as the actual treatment month. We then estimate the TWFE specification in column (1) of Table A-1. We repeat this procedure 10,000 times to generate a null distribution of the parameter estimate. We then compute a p-value by calculating the proportion of simulated coefficient estimates that are at least as small as the actual observed estimate.

The result of this exercise is presented in Figure A-3, which shows a histogram of the null distribution generated using the placebo approach described above. The vertical line shows the actual estimate reported in column (1) of Table A-1. The estimated one-sided p-value is 0.139. In other words, roughly 13% of simulations generated a point estimate less than -0.296. We interpret this to be weak evidence in favor of the Salah effect hypothesis.

[5] In implementing this method, we omit data from West Yorkshire because we only have data for two pre-treatment months. In all, there are 23 placebo units we use for this procedure.

[6] Data were downloaded from the U.K. police data download page (`https://data.police.uk/data/`).

	(1)	(2)	(3)	(4)
Treated	-0.296^{***}	-0.214^{***}	-0.288^{***}	-0.152^{*}
	(0.0610)	(0.0488)	(0.0903)	(0.0805)
Observations	969	969	969	969
R-squared	0.896	0.932	0.913	0.942
Police Force FE	✓	✓	✓	✓
Month FE	✓	✓	✓	✓
Unit-specific time trend		✓		✓
Weights			✓	✓

Table A-1: Regression results with monthly annualized hate crime rate as the dependent variable. Robust standard errors, clustered by police force, are reported in parentheses. For comparison, the estimated ATT yielded by the matrix completion method, averaging across post-treatment months, was -0.275. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

A.5 Are London and Manchester Driving the Results?

As noted in the section "Robustness and Generalizability Tests", there were two terrorist attacks just prior to Salah joining Liverpool F.C. — one in Manchester and one in London. To confirm that our results are not being driven by an increase in hate crimes in these cities in response to the attacks, we re-estimate the matrix completion model for hate crimes without Manchester and London. The results are virtually identical to those obtained from the full data.

Figure A-4 plots the difference between imputed and observed outcomes for each month using the full data (horizontal axis) against the difference when we omit Manchester and London. Each point is a month in the data. The 45-degree line is also plotted. All points fall very close to the 45-degree line, which indicates that the results are not being driven by Manchester and London.

B Twitter Analysis

B.1 Data Collection

As of 2018, about one quarter of the U.K.'s population was an active Twitter user. While this constitutes a large subsection of the U.K. population, recent research indicates that U.K. Twitter users are not representative of the U.K. population as a whole. They are disproportionately young, male, and more likely to have managerial, administrative, and professional occupations (Sloan, 2017). However, the platform is widely used by British soccer fans, with 3 of the top 20 most followed accounts in the U.K. belonging to English Premier League teams, alongside popular news accounts like the BBC and celebrities such as Harry Styles and Emma Watson (Social Backers, 2019). Twitter data thus gives us access to public messages produced by a large cross-section of U.K. soccer fans.

Looking at soccer fans based in the U.K., we compare the frequency of anti-Muslim tweets published by fans of Liverpool F.C. relative to fans of other English teams over time. We began by using Twitter's API to scrape the account IDs of all followers of the top five most followed teams in the English Premier League: Manchester United F.C. (19 million followers), Arsenal F.C. (14 million),

Accessed on May 22, 2019.

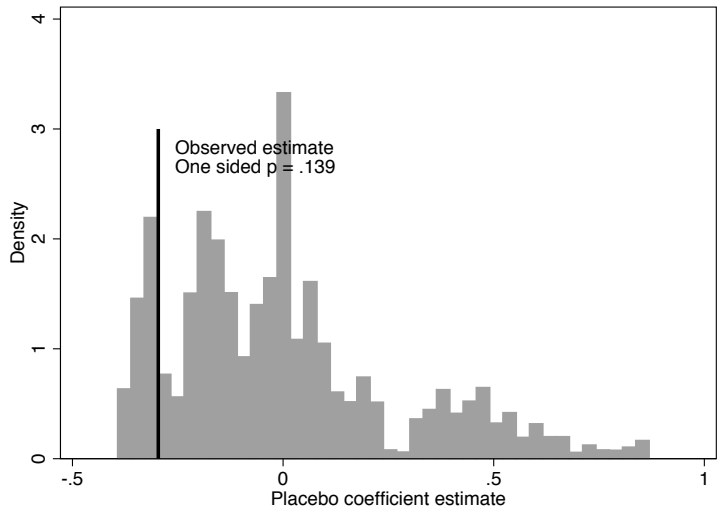


Figure A-3: The histogram shows the simulated null distribution of difference-in-differences estimates. The solid black line shows the observed coefficient for Merseyside. The one-sided p-value is 0.139.

Chelsea F.C. (12 million), Liverpool F.C. (11 million), and Manchester City F.C. (6 million). We also scraped the followers of Everton F.C., a smaller team with 1.75 million followers that is also located in the city of Liverpool. Fans of both clubs are nearly identical in terms of demographics: the home stadiums are within walking distance of each other, there are no historic political, religious, or social differences between their fanbases, and many Liverpudlian families are mixed in their allegiances (Borden, 2014). Evertonians thus constitute the closest comparison group in the sample, with one key difference as a result of their fierce rivalry: exposure to Salah may skew negative for Evertonians, but is positive and goal-aligned for Liverpool F.C. fans.

After obtaining followers' account IDs, we collected our sample of tweets as follows. First, to ensure that the users in our sample had been soccer fans prior to Salah joining Liverpool, we subset our follower IDs to the oldest 500,000 followers of each team. Follower IDs are scraped from Twitter's API in the reverse order that the users began following the account, with newer users appearing first. This feature of the data enables us to identify long-term fans of each team, given that the team accounts have been popular for almost a decade and now have millions of followers. Then, to ensure that users in our sample were located in the U.K., we again used Twitter's API to download profile metadata for the 500,000 oldest followers of each team.[7] We then used their "user.location" metadata field to determine if each user was located in the U.K. based on the text of their self-reported locations.[8] Once

[7]This method ensures that the sample joined Twitter before the treatment. For instance, the 500,000 most recent followers for Liverpool F.C., and the most recent accounts, were created between 2015 and mid-2016.

[8]Users were classified as being located in the U.K. if their "user.location" metadata field contained either a city or country keyword indicating that the user was located in the U.K. City keywords were obtained using the `maps` package in `R`. While this method does not necessarily capture all fans of these soccer teams located in the U.K., as many users do not provide any location metadata at all, it ensures that our sample consists only of likely U.K. residents. As Hecht et al. (2011) demonstrate, a user's country and state can be determined with decent accuracy using self-reported Twitter data, and users often reveal location information with or without realizing it. Similarly, Mislove et al. (2011) explain that because large numbers of users report their location in

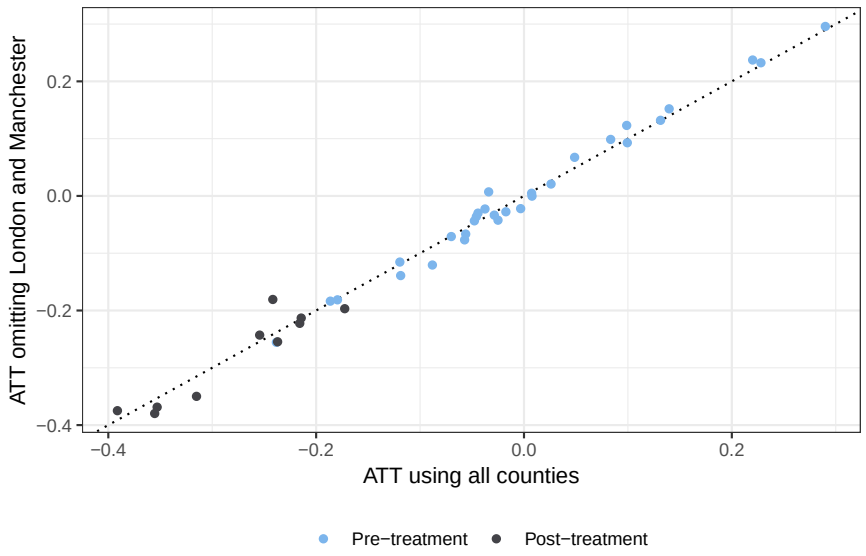


Figure A-4: Difference between observed and imputed outcomes for Merseyside using the full dataset (horizontal axis) and omitting Manchester and London (vertical axis). Each point is a month. The 45-degree line is shown.

we identified longtime Twitter followers of English Premier League teams that were likely to be located in the U.K., we randomly sampled 10,000 followers from each team. We used Twitter's API a final time to scrape up to 3,200 of the most recent tweets published by each of these 60,000 U.K. soccer fans.[9] This resulted in a dataset of approximately 15 million tweets produced by the 60,000 English followers of the "Big Five" clubs of English soccer plus Everton F.C.

In order to identify anti-Muslim tweets, which are relatively rare in this dataset of all tweets produced by soccer fans in the U.K. (approximately .03% of all tweets), we first identified all tweets broadly about Muslims in our dataset. We began with the terms "muslim" and "islam" and used a word2vec model (a neural network that processes text) to find other relevant terms in the data. This yielded the following broad relevant keywords: "arab," "arabs," "islam," "muslims," "muslim," "mosque," and "mosques."[10] About 44,000 of the 15 million tweets in our dataset contained one of these relevant keywords. We then took a sample of about 1,500 of these tweets containing a keyword relevant to Muslims or Islam and used Figure8 (formerly Crowdflower), a crowd-sourced data enrich-

the "user.location" field and in aggregate these reports are quite accurate, this is a reasonable way to determine a user's location. This is particularly true given that we are more interested in obtaining a high degree of precision (ensuring that the users are actually U.K. residents) than recall (obtaining the entire population of tweets sent by U.K. residents).

[9]The 3,200 tweet limit is imposed by Twitter's API and for most Twitter users covers their entire Twitter timelines beginning on the day they first joined the platform.

[10]The word2vec model also identified many irrelevant keywords to our study such as "rohingya" (in reference to the ongoing conflict in Myanmar) and "assad" (in reference to the Syria conflict). We only chose to include relevant keywords that were among the top 50 words that the word2vec model indicated were most similar to the terms "muslim" and "islam." Although most British Muslims are of South Asian descent, the word "pakistani" did not appear in the top 50 words identified by the word2vec model and therefore we did not use it as keyword to filter our data.

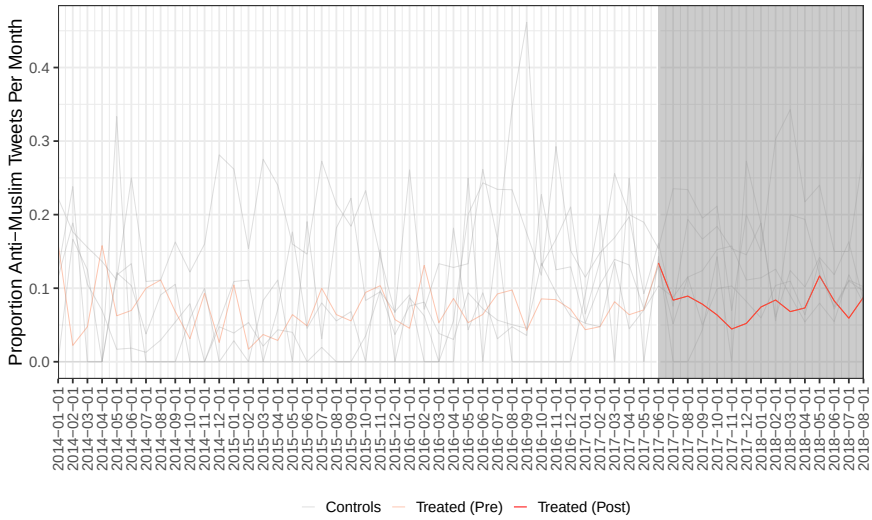


Figure A-5: Anti-Muslim tweets as a proportion of tweets mentioning Muslims or Islam, across followers of British soccer clubs. The red line shows anti-Muslim tweets by followers of Liverpool F.C.

ment platform, to have three native English speakers code each of these 1,500 tweets as anti-Muslim or not.[11]

B.2 Twitter Coding Instructions

The instructions provided to coders on Figure8 (formerly Crowdflower) were as follows:

Overview: In this job, you will be presented with tweets about Muslims and Islam. Review the tweets to determine the sentiment so that we can have a greater understanding about the overall sentiment expressed by the author.

Steps: (1) Read the tweet. (2) Determine if the tweet is relevant to Muslim people or Islam. (3) Determine if the tweet expresses a positive, neutral, or negative attitude towards Muslims or Islam.

Rules & Tips: The posts can be classified as positive, negative or neutral:

- **Positive tweets** portray Muslim people or Islam in a positive manner or argue that Muslims and Islam should not be portrayed negatively. For example, tweets that state that Muslims are not terrorists or extremists or that Islam is a peaceful religion or tweets that defend Muslims or Islam are positive.
- **Neutral tweets** are only informative in nature and provide no hint as to the mood of the author. They do not express an opinion about Muslims or Islam.
- **Negative tweets** are tweets in which some aspects of the tweet uncover a negative mood such as, criticism, insults or a negative comparison. These include tweets portraying Muslims as terrorists, extremists, or violent, and those making negative generalizations about Muslims or Islam as a whole.

[11] The instructions provided to coders are displayed in Appendix B. For more information on using Figure8 (formerly Crowdflower) to code data for training classifiers, see Benoit et al. (2016).

- **Irrelevant tweets** do not mention Muslims or Islam or are not in English. These include tweets where the word "Muslim" or "Islam" appears in the handle of a Twitter user and tweets in foreign languages, for example.

- **Note:** Tweets that are purely factual (links to news articles without comment) are not necessarily Neutral — consider whether the fact/news itself is Positive or Negative regarding the business and select one of those when possible.

B.3 Twitter Data Descriptive Statistics

Table A-2: Proportion of Anti-Muslim Tweets Pre and Post-Salah

type	team	post_salah	mean
Anti-Muslim / Muslim Relevant	liverpool	0	0.073
Anti-Muslim / Muslim Relevant	liverpool	1	0.076
Anti-Muslim / Muslim Relevant	other teams	0	0.102
Anti-Muslim / Muslim Relevant	other teams	1	0.115

B.4 Twitter Data Additional Data Analysis

As an alternative method of analyzing the effect of Salah joining Liverpool on the monthly proportion of anti-Muslim tweets produced by Liverpool fans, we conduct difference-in-differences estimation as follows:

$$y = \beta_0 + \beta_1 T + \beta_2 L + \beta_3 (T \cdot L) + \varepsilon \tag{A-2}$$

Here T is a dummy variable for the time period, equal to 1 in the post-Salah period and 0 in the pre-Salah period, and L is a dummy variable for Liverpool group membership, equal to 1 for Liverpool and 0 for other teams. The interacted term $(T \cdot L)$ is a dummy variable indicating when $L = T = 1$. If the coefficient β_3 on $(T \cdot L)$ is negative, as expected, then Liverpool fans tweet less anti-Muslim content in the post-Salah period relative to the pre-Salah period, compared to fans of other teams. We conduct this analysis comparing Liverpool fans' tweets to tweets produced by fans of other large teams as well as Everton F.C.. We use the proportion of anti-Muslim tweets (anti-Muslim tweets / tweets relevant to Islam or Muslims) as our outcome variable y.

Because there is only one treated unit and standard errors may be misleading, we again undertake an alternative form of inference, whereby we randomly assign a single unit to be treated, with treatment beginning in a randomly chosen month that is at least 4 months after the first observations in our dataset and as late as the actual treatment month. We then estimate the difference-in-difference model above. We repeat this procedure 10,000 times to generate a null distribution of the parameter estimate. We then compute a p-value by calculating the proportion of simulated coefficient estimates that are at least as small as the actual observed estimate.

The result of this exercise is presented in Figure A-6, which shows a histogram of the null distribution generated using the placebo approach described above. The vertical line shows the actual estimate of the model in equation A-2. The estimated one-sided p-value is 0.07. In other words, roughly 7% of simulations generated a point estimate less than -0.038. We interpret this to be suggestive evidence in favor of our Salah effect hypothesis.

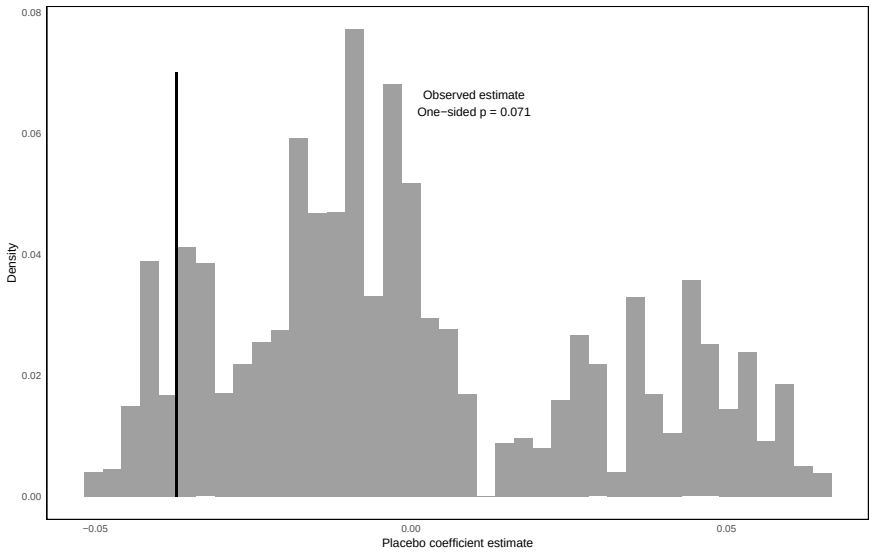


Figure A-6: The histogram shows the simulated null distribution of difference-in-differences estimates. The solid black line shows the observed coefficient for Liverpool. The one-sided p-value is 0.07.

B.5 Testing for Backlash

In order to increase our confidence that we are indeed measuring a decrease in the use of anti-Muslim discourse by Liverpool fans, relative to fans of rival teams, we conduct a difference in differences analysis comparing tweets of rival team fans to tweets of Twitter users located in the UK who do not follow any soccer teams. If our results are driven by backlash from rival team fans, we would expect to see an increase in anti-Muslim discourse from these fans after Salah joins Liverpool, relative to non-soccer fans in the UK.

To identify these non-soccer fans located in the UK we first collect 30,000 recent tweets from Twitter users who are geolocated in the UK. We then filter these accounts to individuals who do not follow any of the major football team accounts, who had active Twitter accounts during our period of analysis, leaving us with a sample of about 15,000 unique users. We then scrape the 3200 most recent tweets from each of these accounts and use our same method to classify their tweets.

The results of our differences in differences analysis, reported in Table A-3 suggest that there is no "Salah backlash effect" in which fans of rival teams publicly express more anti-Muslim sentiment after Salah joins Liverpool, relative to non-soccer fans in the UK. This increases our confidence that we are actually measuring a decrease in anti-Muslim Tweets by Liverpool fans, rather than a change driven by backlash from rival teams.

Table A-3: Effect of Salah Joining Liverpool on Daily Proportion of Anti-Muslim Tweets

	Proportion of Anti-Muslim Tweets
Constant	0.139^{***}
	(0.012)
Non-Liverpool Fans (Treated Dummy)	-0.045^{**}
	(0.014)
Post-Salah (Post-Treatment Dummy)	-0.014
	(0.035)
Non-Liverpool Fans x Post-Salah (DID)	0.047
	(0.039)
R^2	0.047
Adj. R^2	0.037
Num. obs.	288
RMSE	0.081

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

C Mané Effect Analysis

C.1 Liverpool Echo

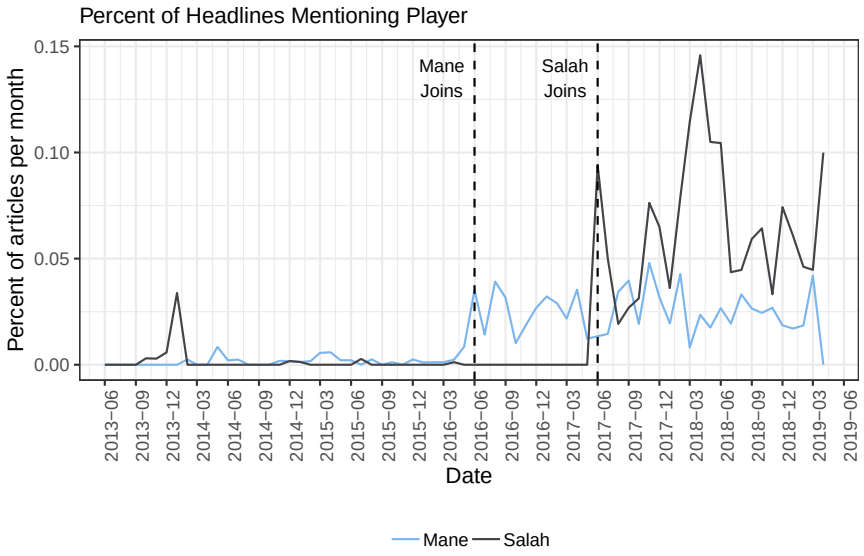


Figure A-7: Percent of monthly titles in Liverpool Echo that mention Mané or Salah

C.2 Hate Crimes

Here, we repeat the same matrix completion analysis of hate crime data as in the main text, except treating July 2016 — the month in which Sadio Mané signed with Liverpool — as the beginning of treatment. Additionally, to avoid picking up the Salah effect, we truncate the data to before Salah signed.

The results are shown in Figure A-8. Overall, we see no consistent difference between observed and imputed hate crimes in Merseyside after Mané joined Liverpool (but before Salah joined). Averaging

across post-treatment months, the estimated ATT is 0.017 (S.E. $= 0.049$), which corresponds to a 1.3% *increase* in the hate crime rate — though this result is not statistically significant.

C.3 Twitter

We also repeat the same matrix completion analysis of Twitter data as in the main text, except treating July 2016 — the month in which Sadio Mané signed with Liverpool — as the beginning of treatment. Additionally, to avoid picking up the Salah effect, we truncate the data to before Salah signed.

The results are shown in Figure A-9. Unlike the hate crime data, here we do observe a significant decrease between observed and imputed anti-Muslim tweets in Merseyside after Mané joined Liverpool (but before Salah joined). Averaging across post-treatment months, the estimated ATT is -0.043 (S.E. $= 0.007$), which corresponds to a 59.8% *decrease* in the proportion of anti-Muslim tweets.

D Survey Experiment

D.1 Survey experiment design

The survey experiment is a 2×2 factorial design embedded in the survey, in addition to a pure control group. First, we provided the treated respondents with a vignette emphasizing Salah's success (*Success Condition*) or speculation about his potentially declining performance (*Failure Condition*).[12] This factor is designed to test the "model minority" dimension of the positivity condition for prejudice reduction. Next, treated respondents saw another vignette emphasizing either Salah's religiosity (*Religiosity Condition*) or agreeable character (*Character Condition*).

D.2 Vignette Descriptions

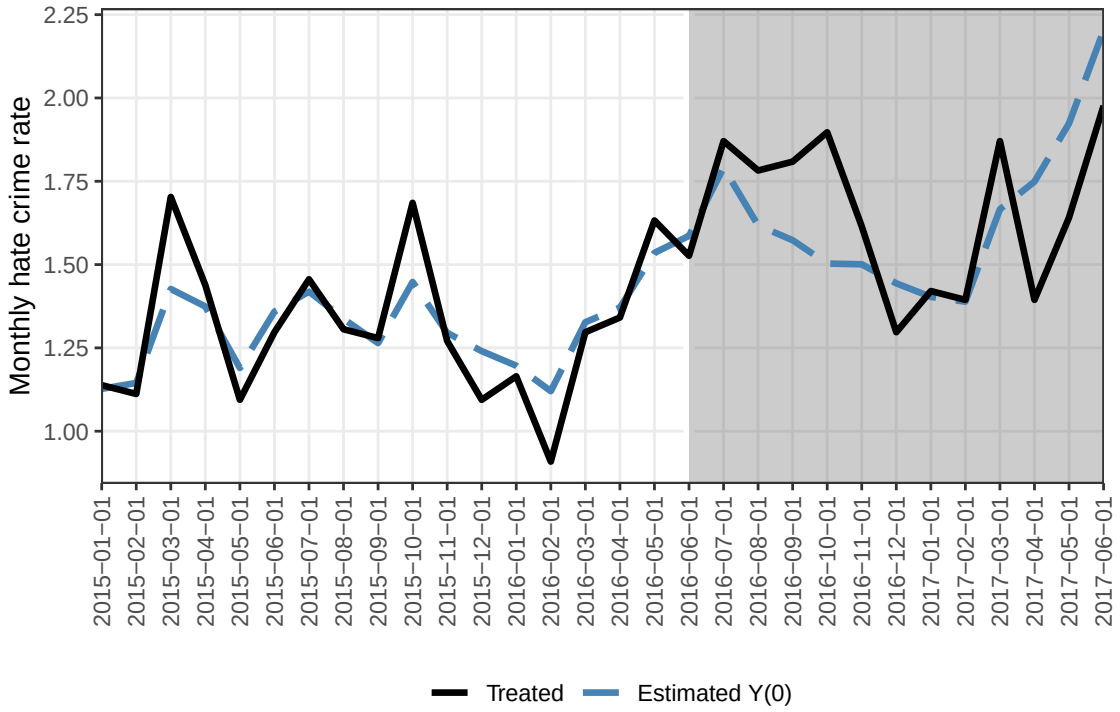
Respondents in the success condition saw a picture of Mo Salah holding the Golden Boot with the following text:

> *In the 2017-18 season, Salah scored 43 goals for Liverpool, setting numerous club and league records along the way. For his efforts, he was named the Premier League's Player of the Month three times, won the Golden Boot, and was awarded the PFA Players' Player of the Year award. Along with Christiano Ronaldo and Luca Modric, he was shortlisted for UEFA Player of the Year. He recently won the FIFA Puskás award for best goal.*
>
> *Salah was also central in taking Egypt to the World Cup and Liverpool F.C. to the Champions League final.*

Respondents in the failure treatment saw an image of Salah looking regretful with the following text:

[12] In the beginning of the 2018-2019 season, when we fielded the experiment, Salah got off to a slower start than the previous season, which the vignette in the *Failure Condition* emphasized.

(a) Observed and imputed outcomes for Merseyside



(b) Estimated ATT in every period

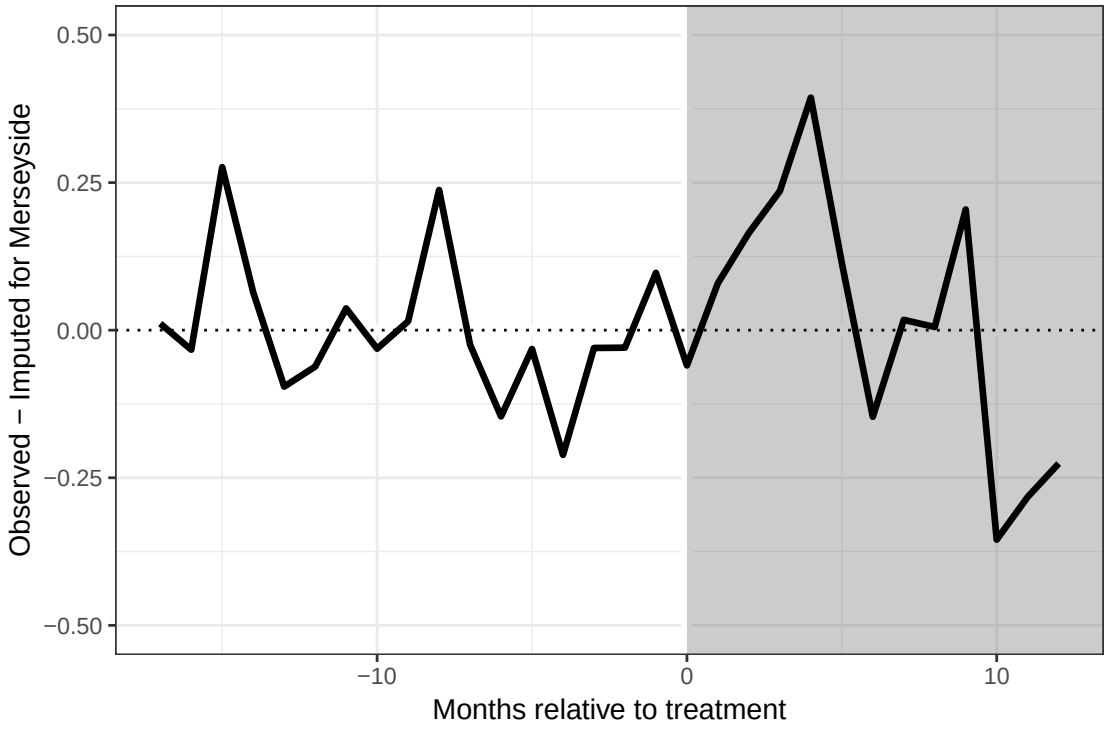
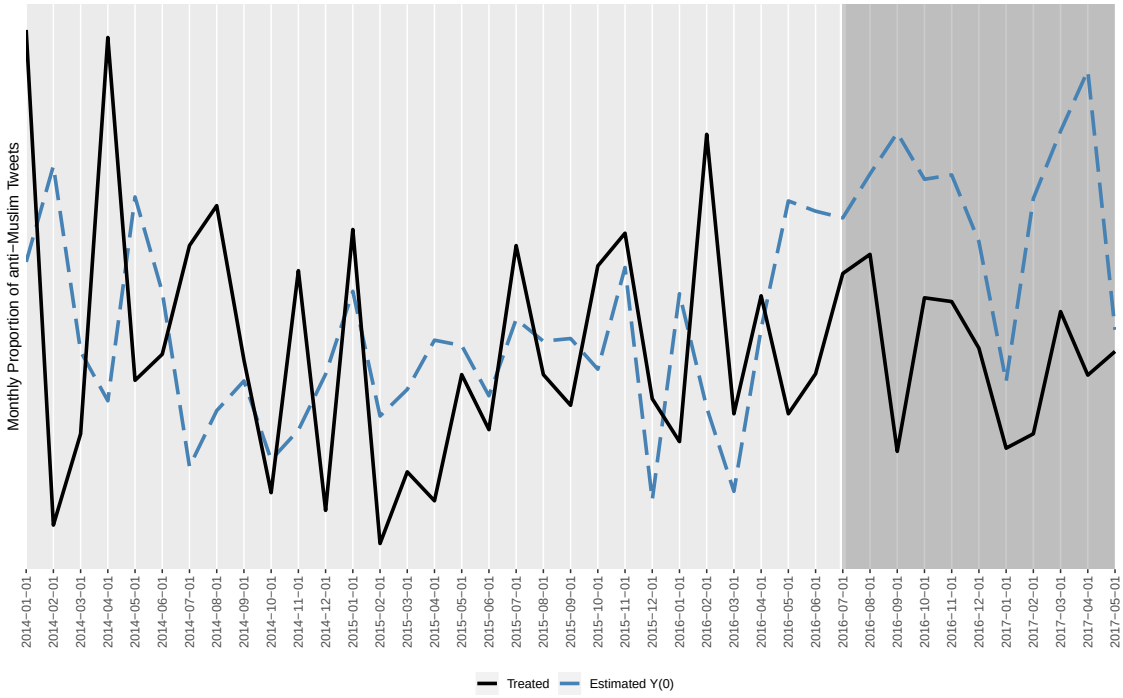


Figure A-8: Matrix completion results for hate crime outcomes, treating Sadio Mané's signing as the beginning of treatment. The top panel shows the observed (solid line) and imputed (dashed line) monthly hate crime rates in Merseyside. The bottom panel shows the difference between the observed and imputed outcomes. In the post-treatment period, this is the estimate of the ATT.

(a) Observed and imputed outcomes for Merseyside



(b) Estimated ATT in every period

Figure A-9: Matrix completion results for tweet outcomes, treating Sadio Mané's signing as the beginning of treatment. The top panel shows the observed (solid line) and imputed (dashed line) monthly proportion of anti-Muslim tweets produced by Liverpool fans. The bottom panel shows the difference between the observed and imputed outcomes in Liverpool fans' tweets (red line) relative to tweets produced by fans of other U.K. football clubs (gray lines). In the post-treatment period, this is the estimate of the ATT.

Despite a successful 2017-18 season, some believe he is underperforming this season. As of late October, he had scored only 4 goals in Premier League play.[13] *Due to this lackluster performance, some critics have suggested that Salah will be a 'one-season wonder.'*

After the success/failure treatment, respondents then received a treatment emphasizing either Salah's character or his religiosity. Respondents who received the character treatment saw a picture of Salah with his daughter and the following text:

In addition to his goal-scoring, Salah is known for his character both on and off the pitch. In his native Egypt, Salah privately donated millions of pounds to charity and to a leading anti-drug campaign. Always a sportsman, Salah does not celebrate goals against his former teams and picked up only two yellow cards in 49 matches for Liverpool last season.

Respondents in the religious treatment saw Salah prostrating with this text:

In addition to his goal scoring, Salah is known for an attachment to his Muslim identity both on and off the pitch. After every goal he scores, Salah touches his head to the ground in prayer. He also fasts during Ramadan (except on match days) and shares well wishes with his followers on social media during Islamic holidays. He named his daughter Makka after Islam's holiest site (Mecca).

D.3 Balance Table for Survey Experiment

Table A-4 shows a balance table across arms of the survey experiment.

D.4 Heterogeneous Treatment Effects

We examine heterogenous treatment effects across social and political characteristics. Table A-7 shows that the religion prime has a similar effect regardless of self-identified political ideology. Table A-8 shows the results of a Lin (2013) regression that interacts the religion treatment with demeaned demographic covariates. There are no significant interactions by age, sex, or education. Table A-9 estimates heterogeneous effects according to whether the respondent resides in Liverpool. There are positive interactions between living in Liverpool and the treatment, but they are imprecisely estimated. A larger set of heterogeneous effects are reported in the full appendix, posted on Dataverse at `https://doi.org/10.7910/DVN/2JKWNS`.

[13] This statistic was updated for respondents taking the survey in or after January 1, 2019 to read: "As of early January, he had scored in just 62% of Premier League games played — compared to 89% last season.

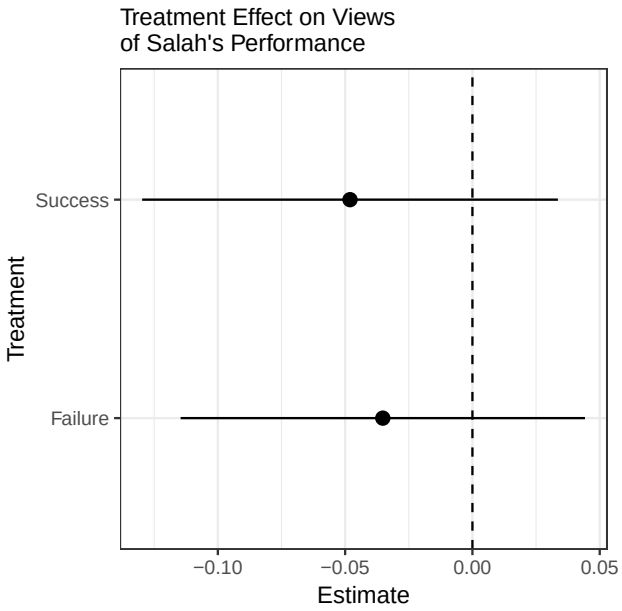


Figure A-10: Effects of treatments on views of Salah's performance

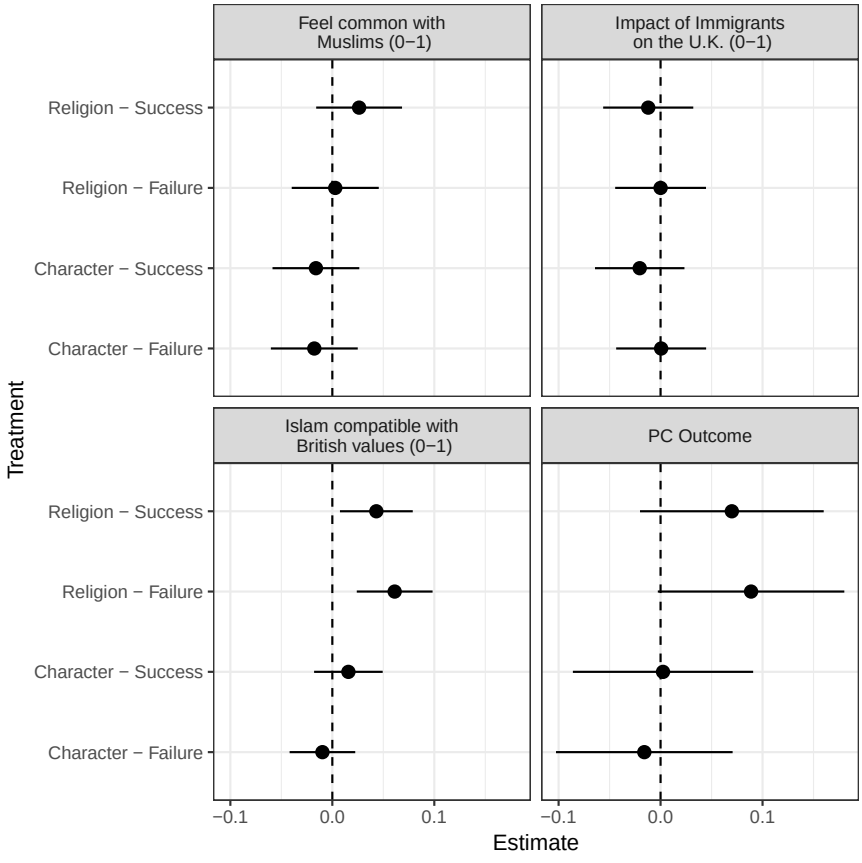


Figure A-11: Coefficient plots representing the average treatment effects on the four outcomes, relative to the pure control condition. The first three outcome variables are binary, while the the fourth is a continuous variable with mean of zero and unit variance. Bars show 95% robust confidence intervals.

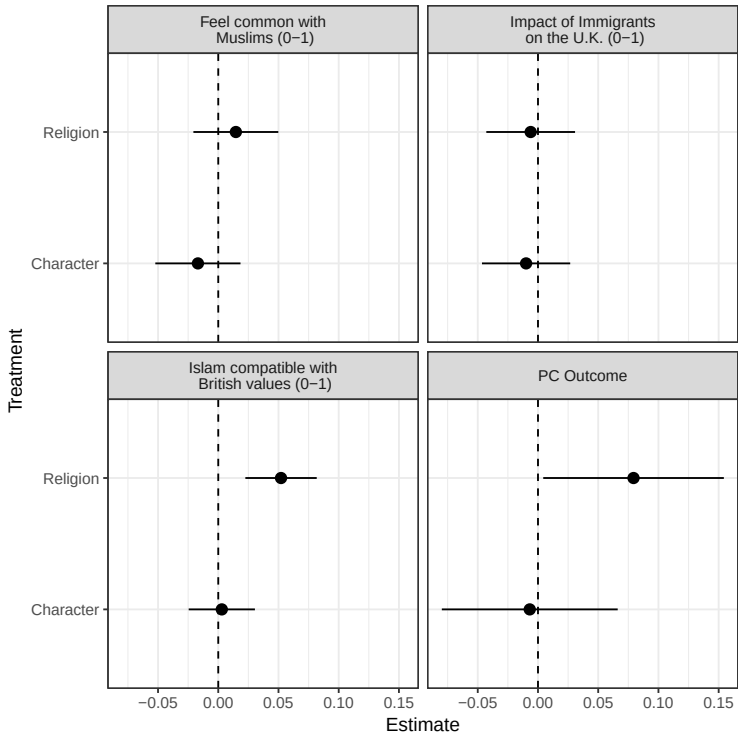


Figure A-12: Coefficient Plot for the average marginal component effects of the religion/character treatment

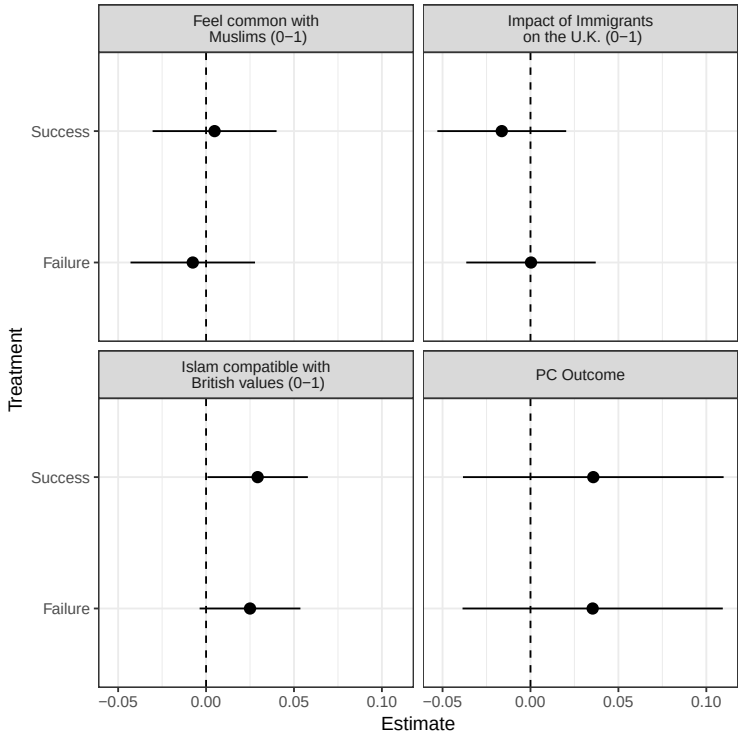


Figure A-13: Coefficient Plot for the average marginal component effects of the success/failure treatment

	Control (N=2887)	Char. - Fail. (N=1463)	Char. - Succ. (N=1454)	Rlgn. - Fail. (N=1421)	Rlgn - Succ. (N=1441)	F-Stat (p.value)
Age (Years)						0.89 (0.47)
N-Miss	136	161	170	204	173	
Mean (SD)	49.90 (12.84)	50.46 (12.22)	49.90 (12.71)	50.23 (12.43)	49.78 (11.44)	
Range	18.00 - 98.00	18.00 - 98.00	18.00 - 98.00	18.00 - 98.00	18.00 - 98.00	
Female						1.95 (0.1)
N-Miss	1	20	0	5	0	
Mean (SD)	0.28 (0.45)	0.27 (0.44)	0.25 (0.44)	0.25 (0.43)	0.28 (0.45)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	
University Edu.						1.66 (0.16)
N-Miss	6	18	5	38	14	
Mean (SD)	0.32 (0.47)	0.34 (0.48)	0.31 (0.46)	0.33 (0.47)	0.31 (0.46)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	
Salah Favorite						0.02 (1)
N-Miss	560	686	466	615	413	
Mean (SD)	0.52 (0.50)	0.52 (0.50)	0.52 (0.50)	0.53 (0.50)	0.52 (0.50)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	
Karius Empathy						0.5 (0.73)
N-Miss	217	381	274	333	244	
Mean (SD)	0.38 (0.48)	0.36 (0.48)	0.38 (0.49)	0.39 (0.49)	0.38 (0.49)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	
Liverpool Resident						0.75 (0.56)
N-Miss	0	0	0	0	0	
Mean (SD)	0.22 (0.42)	0.24 (0.43)	0.23 (0.42)	0.23 (0.42)	0.22 (0.41)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	
Conservative						0.24 (0.91)
N-Miss	241	188	193	171	233	
Mean (SD)	0.27 (0.44)	0.27 (0.44)	0.28 (0.45)	0.28 (0.45)	0.28 (0.45)	
Range	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	0.00 - 1.00	

Table A-4: Summary statistics for several demographic variables and other characteristics by treatment group. *Female* indicates proportion of respondents who identified as females. *University Edu.* indicates proportion of respondents who have at least some university education. *Salah Favorite* indicates the proportion of respondents who indicated that Salah is their favorite player. *Karius Empathy* indicates those who expressed empathy with Liverpool's goalkeeper Karius. *Liverpool Resident* indicates whether the respondents live in Liverpool. *Conservative* indicates respondents who indicated they are associated with the Conservative or UK Independence Party.

	PC Outcome	Muslims Common	Islam Compatible	Immigrant Impact
Constant	-0.02	0.57^{***}	0.18^{***}	0.46^{***}
	(0.03)	(0.01)	(0.01)	(0.01)
Religion	0.08^{*}	0.01	0.05^{***}	-0.01
	(0.04)	(0.02)	(0.02)	(0.02)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	0.00	0.00	0.00	-0.00
Num. obs.	4997	5361	5168	5032
RMSE	2.14	1.05	0.87	1.06

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table A-5: Main regressions using the character/religion treatments.

	PC Outcome	Muslims Common	Islam Compatible	Immigrant Impact
Constant	-0.02	0.57^{***}	0.18^{***}	0.46^{***}
	(0.03)	(0.01)	(0.01)	(0.01)
Failure	0.04	-0.01	0.02	0.00
	(0.04)	(0.02)	(0.01)	(0.02)
Success	0.04	0.00	0.03^{*}	-0.02
	(0.04)	(0.02)	(0.01)	(0.02)
R^2	0.00	0.00	0.00	0.00
Adj. R^2	-0.00	-0.00	0.00	0.00
Num. obs.	7515	8060	7771	7571
RMSE	2.26	1.11	0.90	1.12

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table A-6: Main regressions using the success/failure treatments.

	PC Outcome	Muslims Common	Islam Compatible	Immigrant Impact
(Intercept)	0.08^{*}	0.61^{***}	0.21^{***}	0.51^{***}
	(0.03)	(0.02)	(0.01)	(0.02)
Character	-0.02	-0.03	-0.00	-0.01
	(0.05)	(0.02)	(0.02)	(0.02)
Religion	0.09	0.02	0.05^{**}	-0.01
	(0.05)	(0.02)	(0.02)	(0.02)
Conservative	-0.38^{***}	-0.14^{***}	-0.12^{***}	-0.15^{***}
	(0.06)	(0.03)	(0.02)	(0.03)
Character:Conservative	0.06	0.04	0.02	0.01
	(0.08)	(0.04)	(0.03)	(0.04)
Religion:Conservative	-0.03	-0.03	-0.00	0.00
	(0.08)	(0.04)	(0.03)	(0.04)
R^2	0.03	0.02	0.02	0.02
Adj. R^2	0.03	0.02	0.02	0.02
Num. obs.	7372	7900	7617	7417
RMSE	2.22	1.10	0.89	1.11

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table A-7: Interacting the character/religion treatments with an indicator for conservative views. This indicator is coded as 1 if the respondent identifies with the Conservative Party or the UK Independence Party. It is coded as 0 if the respondent identifies with the Labour Party, Liberal Democrats, other parties, or none of these parties.

	PC Outcome	Muslims Common	Islam Compatible	Immigrant Impact
(Intercept)	-0.02	0.57^{***}	0.18^{***}	0.47^{***}
	(0.03)	(0.01)	(0.01)	(0.01)
Character	-0.00	-0.01	0.00	-0.01
	(0.04)	(0.02)	(0.01)	(0.02)
Religion	0.08^{*}	0.02	0.05^{***}	-0.00
	(0.04)	(0.02)	(0.01)	(0.02)
Age	0.00	-0.00^{**}	0.00^{*}	0.00^{*}
	(0.00)	(0.00)	(0.00)	(0.00)
Female	0.06	0.06^{*}	-0.00	-0.00
	(0.06)	(0.03)	(0.02)	(0.03)
Univ. Edu.	0.68^{***}	0.24^{***}	0.22^{***}	0.25^{***}
	(0.06)	(0.03)	(0.03)	(0.03)
Character:Age	-0.00	0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Religion:Age	-0.00	0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Character:Female	0.01	-0.01	0.02	-0.00
	(0.08)	(0.04)	(0.03)	(0.04)
Religion:Female	0.08	0.01	0.03	0.04
	(0.08)	(0.04)	(0.03)	(0.04)
Character:Univ. Edu.	0.03	0.00	-0.04	0.08^{*}
	(0.08)	(0.04)	(0.03)	(0.04)
Religion:Univ. Edu.	-0.03	-0.03	-0.01	0.01
	(0.08)	(0.04)	(0.04)	(0.04)
R^2	0.11	0.06	0.06	0.08
Adj. R^2	0.11	0.06	0.06	0.08
Num. obs.	7377	7914	7627	7429
RMSE	2.13	1.08	0.87	1.08

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table A-8: Lin regressions using the character/religion treatments.

	PC Outcome	Muslims Common	Islam Compatible	Immigrant Impact
(Intercept)	-0.06	0.57^{***}	0.17^{***}	0.45^{***}
	(0.03)	(0.02)	(0.01)	(0.02)
Character	-0.01	-0.02	0.01	-0.01
	(0.04)	(0.02)	(0.02)	(0.02)
Religion	0.05	0.01	0.05^{**}	-0.02
	(0.04)	(0.02)	(0.02)	(0.02)
Liverpool Res.	0.15^{*}	0.02	0.06^{*}	0.07^{*}
	(0.07)	(0.03)	(0.03)	(0.03)
Character:Liverpool Res.	0.03	0.02	-0.02	0.02
	(0.09)	(0.04)	(0.04)	(0.04)
Religion:Liverpool Res.	0.14	0.04	0.03	0.06
	(0.10)	(0.04)	(0.04)	(0.05)
R^2	0.01	0.00	0.01	0.01
Adj. R^2	0.01	0.00	0.01	0.01
Num. obs.	7515	8060	7771	7571
RMSE	2.25	1.11	0.89	1.12

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table A-9: Interacting the character/religion treatments with an indicator for residing in Liverpool.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies : Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens and Khashayar Khosravi. 2018. "Matrix completion methods for causal panel data models.".

BBC News. 2018. "Religious hate crimes: Rise in offences recorded by police." *BBC News* .
URL: *https://www.independent.co.uk/sport/football/premier-league/kashif-siddiqi-interview-premier-league-kashif-siddiqi-foundation-altus-league-a8283736.html*

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* 110(2):278–295.

Borden, Sam. 2014. "The Friendly Derby? Well, Everton-Liverpool Is Friendlier." *The New York Times* .
URL: *https://www.nytimes.com/2014/01/28/sports/soccer/the-friendly-derby-well-everton-liverpool-is-friendlier.html*

Doudchenko, Nikolay and Guido W. Imbens. 2016. "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis." *National Bureau of Economic Research* (No. w22791).
URL: *https://www-nber-org.stanford.idm.oclc.org/papers/w22791.pdf*

Hecht, Brent, Lichan Hong, Bongwon Suh and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM pp. 237–246.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Annals of Applied Statistics* 7(1):295–318.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

Sloan, Luke. 2017. "Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015." *Social Media+ Society* 3(1):1–11.

Social Backers. 2019. United Kingdom Social Marketing Reports. Technical report.
URL: *https://www.socialbakers.com/statistics/twitter/profiles/united-kingdom/*

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25:57–76.