Overcoming the Trust Deficit: A Field Experiment on Inter-Group Contact in Iraq

Pre-Analysis Plan Update

Salma Mousa*

November 8, 2018

Contents

*PhD Candidate, Political Science, Stanford University. smousa@stanford.edu.

Building on the results of the pilot experiment fielded in 2017 (PAP #20170603AA), I will launch a scale-up of the RCT starting in September 2018. The scale-up includes a larger sample size, the addition of a comparison group, and an expanded range of outcomes. The experimental protocol in the original PAP holds unless otherwise noted. This document serves three purposes: (1) to outline the randomization process, (2) to describe the additional set of outcomes, and (3) to create the t1 attitudinal indices based on a sample of t0 surveys.

1 Research Design

There are around 50 amateur male soccer teams in the Ankawa and Qaraqosh neighborhoods in the greater Erbil area, almost exclusively segregated by religion. I have randomly selected 37 Christian teams from this set to participate in the tournament, all of which accepted. The 37 teams are divided into three concurrent leagues, two of which are experimental. The two experimental leagues ('Seminaire" and "Qaraqosh Sports Club") consist of 14 teams each. Seven of these teams are assigned to the treatment group (receiving additional Muslim players) while the other seven are assigned to control group (receiving additional Christian players). I block on the a primary attitudinal outcome, a survey item on empathy toward Muslims, before randomizing.[1]

The nature of sports tournaments is such that that even those on control teams will come into contact with Muslims. Control participants will encounter Muslims as opponents and in the wider league in environment, such as during prize-giving ceremonies and in between training sessions. To get a sense of the independent effect of participating in a league, the third league ("Ankawa") serves as a comparison group. The nine teams in the comparison league are all-Christian, and each absorb three additional Christian players. There are thus no Muslims involved in the comparison league. I will show that the players in the comparison league share similar baseline characteristics to those in the two experimental teams, although assignment to this league is based on geography.[2] With twelve players per team (plus one coach), and 37 teams in total, the sample size is thus $n = 481$.[3] When pooled with pilot results, the sample grows to 51 teams and 649 participants.

The teams combine players across a range of ages (16 – 43) and were founded at roughly the same time, shortly after the mass displacement of August 6, 2014. The captains were told that a well-known local NGO[4] was setting up a soccer tournament for displaced people in the area, with two conditions for participating. First, that each team will receive an additional three players that may or may not be fellow Christians. These three players join with each team of seven for a total of ten players per team. Second, all players have to take a brief survey on the displacement experience and their views on Iraqi society before and after the league.

In addition to recruiting existing amateur Christian *teams*, I recruit 42 Muslim and 69 Christian individual soccer *players*. I randomly chose these individuals from team rosters for amateur Muslim and Christian teams that were not randomly chosen to participate in the tournament. Selecting

[1] The question reads: "How much do you agree with this statement: I have a lot in common with Sunni Arabs." I divide the t0 responses into seven blocks before randomizing which paired team will receive the treatment.

[2] Participation in the comparison league is based on residing closer to the field in the Ankawa neighborhood than the Qaraqosh Sports Club or Seminaire fields, about a 40 minute drive away.

[3] Each team also has one (Christian) coach, whose outcomes I also measure as a secondary analysis.

[4] MaakThaTheh (*"together"* in Assyrian Neo-Aramaic), is an Iraqi Christian-led NGO serving IDPs and the operational partner for this study.

Muslims and Christians who already play in their respective amateur leagues helps to ensure that these added players differ only on religious identity and do not systematically differ on skill or motivation. I check for balance in skill among added players using enumerator-coded skill levels (from 0 to 10). The 111 contacted players were told that they were selected to participate in a soccer tournament in their neighborhood, and would be added to pre-existing teams.

1.1 Hypotheses

I expect that playing soccer on the same team as Muslims will increase tolerance and trust toward Muslims (those in encountered in the league, and Muslims in Iraq more broadly), and decrease in-group bias toward fellow Christians. I map tolerance, trust, and in-group bias onto experimental outcomes below.

Relative to those on homogenous teams, I hypothesize that Christians assigned to mixed teams are more likely to:

Tolerance

1. Attend a social event open to Muslims in the league and their families one to two weeks after the league ends.
2. Report in the t1 survey that they would not mind joining a mixed team next season.
3. Continue training with Muslims three months after the tournament ends.
4. Patronize Muslim-owned businesses, and spend more at those businesses conditional on attendance, up to three months after the league ends.
5. Report tolerant attitudes, as measured by the t1 survey indices (within a week of the league's end, and again at the one-year mark).

Trust

1. Report in the t1 survey that they would trust a Muslim-owned cash agency with a wire transfer.
2. Conditional on attending the social event, bringing their female relatives to the event.

In-Group Bias

1. Vote for Muslim winners of the Best New Player award in the t1 survey.
2. Choose to donate their t1 survey compensation to an NGO that serves all communities, rather than their own.
3. Have household members that report tolerant attitudes, as measured by the t1 survey.

I also use t-tests to compare differences in means for Muslim players at t0 and t1.

2 Behavioral Outcomes

The original PAP describes the attitudinal outcomes and three key behavioral outcomes: attending a social event open to Muslims three weeks after the league (and bringing one's family conditional on attendance), training with Muslim players encountered within and outside the league four months after the league ends, and adding one's name for consideration for mixed soccer teams in the future. I add four new outcomes: (1) Players' households patronizing Muslim (Christian)–owned businesses, and the money spent conditional on attending, (2) female household members signing up for a mixed community program, (3) player's vote for the best new player, and (4) a survey item asking respondents to choose between donating to a charity that benefits only Christians, only Muslims, or both communities ("We will donate $1 to a charity that you choose. Which charity should we donate to on your behalf?"). I describe each added outcome in detail below.

2.1 Patronage of Outgroup Businesses

As explained in Enos (2017), social geography can limit the long-term effects of contact interventions: "as long as we remain residentially segregated, our current policies for encouraging contact may not be enough, even for those people the policies actually reach" (p. 249). Can contact interventions overcome segregation? I test this proposition in two ways. First, I ask participants at t0 and t1 whether they feel comfortable visiting different neighborhoods or whether they prefer to "stay in [their] own neighborhood." I expect that players on diverse teams – and their households – are more likely to visit mixed or non-coethnic parts of town, relative to those on all-Christian teams.

Second, I institute a voucher system to track whether treated participants are more likely to patronize businesses in non-coethnic parts of town. All players get a voucher for two restaurants: (1) an $8 voucher for restaurant in Muslim-majority Mosul[5] around 35 minutes away by car, and (2) a $5 voucher for a Christian-owned restaurant in Qaraqosh, where the leagues take place. Each voucher is stamped with the player's unique ID and is valid for three months. The restaurant owners share information on money spent, and which vouchers were presented together. Money spent is a rough proxy for how long the respondents stayed in the restaurant.[6]

I expect that Christians on mixed teams are more likely to visit Muslim-owned restaurants, and conditional on visiting, to spend more money than those with homogenous teammates. Tracking which vouchers were used together also tells us whether some teams preferred to travel as a large group, perhaps with the added "protection" of teammates who share the restaurant owner's ethnicity. I expect that players who travel to Mosul – the most challenging restaurant locale both psychologically and geographically – are more likely to travel in larger groups relative to the other two restaurants. I will also interview the Christian restaurant owners to understand if and how they have adapted their sectarian entrance policies in light of the uptick in Muslim customers.[7]

[5]Interviews with 17 Christian returnees to Qaraqosh in September 2018 revealed that none had visited Mosul since the ISIS occupation out of distrust toward Muslims, despite Mosul being a proximate, major city where many had attended university or worked previously.

[6]For example, those who purchase a coffee will likely spend an hour or so at the restaurant, while those who order dinner or shisha will likely spend a few hours there. I take time spent at a restaurant to reflect comfort in a non-coethnic environment.

[7]Cafes, restaurants, and bars owned by Christians in Erbil and surrounding towns enforce strict rules on the admission of Muslim patrons. Muslim customers are only admitted if they are part of a larger group that includes

2.2 Community Program for Female Household Members

Female household members are encouraged to inter-mingle throughout the league by providing a children's play area next to each field, adequate seating, and refreshments. I administer a short survey to female household members, asking if them to choose or posit a community activity they would be interested in (such as a volleyball tournament, drama courses, or community garden) and whether they would prefer that participants in this activity are exclusively Christians. Selecting "I do not care which community the participants are from," and attending the social event with their male relatives, is coded as a tolerant outcome.

2.3 In-Group Bias Vote

At the end of the league, players are asked to vote for the "best new player" in the league (from the pool of added Muslim and Christian players) to receive a prize. Players are not allowed to vote for their own teammates. The research team makes clear that the prize is based entirely on sportsmanship and not skill. Following Lowe (2017), this game measures in-group bias by compelling participants to choose between in-group or out-group members on the basis of preference alone. I expect that players on mixed teams are more likely to choose Muslim players to receive the sportsman award on other teams, relative to players on all-Christian teams.

3 Spillover Surveys

Lasting, community-level change likely requires that contact interventions affect more than just direct participants. To sustain any positive effects, broader norm shifts are needed. I propose expanding the beneficiaries of contact interventions to include local residents. In the norm cascade theory of change I outline in the paper, observing positive inter-group contact will gradually erode the social stigma around interacting with the "other." I investigate the effects of league exposure for non-players in two ways. First, I survey a sample of players' household members after the league. I expect more tolerant attitudes among the wives, siblings, and parents of treated players. I expect this effect to be amplified among those who live closest to the field, as they receive a stronger dose of the treatment.

Second, I distribute an invitation to the final match – along with a coupon for the concession stand – to randomly selected Christian and Muslim households in Qaraqosh. Members of these households will be given a brief (roughly three weeks) before and after the final match. The outcome of interest is an item on secularism: "To what extent do you agree with this statement: it is arbitrary to divide Iraqis in to ethnic and sectarian identities." I expect a tolerant shift at t1 as opposed to t0. Surveying household members and local residents speaks to a pathway for contact interventions to expand their impact beyond direct participants. Expanding the net of beneficiaries gives us more confidence that contact interventions can change local-level social geography.

Christians. Even then, some high-end venues refuse admission to women in headscarves, even if accompanied by Christians. I speculate that Christian business owners may relax their policy of allowing "unaccompanied" Muslims to frequent their restaurant at the end of the voucher period.

4 Attitudinal Outcomes

I add three new survey items: (1) an empathy item ("How much do you agree with this statement: I have a lot in common with Sunni Arabs"), (2) a social geography item ("Do you feel comfortable going to different neighborhoods in the area?"), and a trust item asking respondents to select which ethno-religious groups they would trust to facilitate a cash transfer ("Imagine that you need to send money to a friend in another city. Which of the cash office owners below would you trust? Choose all that apply"). The response options, displayed in a random order, are: "Ahmed, Baghdad", "Mohammed, Erbil", "George, Beirut", and "Behnam, Erbil." These options map on to trust along two dimensions, city and religion, explained in the matrix for Christian respondents below:

City	Religion	Response Option
Same	Same	*Behnam, Erbil*
Same	Different	*Mohammed, Erbil*
Different	Same	*George, Beirut*
Different	Different	*Ahmed, Baghdad*

I expect that treated Christians are more likely to select Ahmed and Mohammed from this set at t1.

For Muslim respondents, the matrix looks like this:

City	Religion	Response Option
Same	Same	*Ali, Erbil*
Same	Different	*Behnam, Erbil*
Different	Same	*Mohammed, Tunis*
Different	Different	*George, Baghdad*

I expect that Muslims are more likely to select Behnam and George from this set at t1 compared with t0.

4.1 Hierarchical Clustering of Outcome Variables

At the time of this writing, 260 players responded to the t0 survey. I take pre-survey responses as covariates to increase statistical precision and collapse individual survey outcomes into more stable indices to extract the most power out of the t0 survey. I use an unsupervised hierarchical machine learning algorithm to identify latent clusters in the pre-survey data. As I show below, these clusters align closely with theoretical expectations.

The following are the attitudinal outcomes of interest:

```
outcome_variables <- c("empathy_t0", "land_t0", "welc_t0", "op_isis_t0",
                       "friends_t0", "areas_t0",
                        "danger_t0", "nbr_sa_t0",
                       "secular_t0", "secular1_t0", "suff_sa_t0",
                       "suff_sh_t0", "nbr_shabak_su",
                       "nbr_shabak_sh", "proud_iraqi_t0")
```

Below is the code used to generate factor loadings associated with each item using a hierarchical clustering method. The squared factor loadings represent the the portion of each variable's variance that is explained by the factor. Summing the squared factor loadings and dividing by the number of variables yields the portion of variance explained by each factor. Using this method, the portion of variance in all variables explained by each factor ranges from around 50% to 72%.

```
# Ensure package recognizes the variables as "qualitative" (ie, categorical)
df <- df %>% mutate_at(vars(one_of(outcome_variables)), funs(as.factor(.)))

# Apply clustering algorithm
tree <- hclustvar(X.quali = df[outcome_variables])
plot.hclustvar(tree, type = "tree")
plot.hclustvar(tree, type = "index")

# Cut the tree at 5 clusters (based on dendrogram)

p5 <- cutreevar(tree, 5)
clusters <- p5$cluster
p5$var


$cluster1 # Coexistence
          squared loading
welc_t0           0.5113161
empathy_t0        0.4501240
proud_iraqi_t0    0.4042764
areas_t0          0.3537081


$cluster2 # Prospects for Peace
                     squared loading correlation
secular_t0      0.5379118
secular1_t0     0.4887048
danger_t0       0.4000896
land_t0         0.3273835


$cluster3 # Tolerance
          squared loading correlation
friends_t0     0.6109202
op_isis_t0     0.6109202


$cluster4 # Muslims as Neighbors
             squared loading correlation
nbr_shabak_sh     0.6272204
nbr_shabak_su     0.5838442
nbr_sa_t0         0.5365770
```

```
$cluster5 # Absolving Muslims of ISIS Blame
          squared loading correlation
suff_sh_t0     0.7444225
suff_sa_t0     0.7444225
```

4.2 Cluster Descriptions

Below are the descriptions of each item included in the five indices, all coded in a pro-trust direction. Two items that do not load clearly onto any of these indices will be analyzed separately, in addition to two items that are not included in the t0 survey.[8]

Index #1: Coexistence

- Believe that Sunni Arabs are welcoming toward Christians
- Proud or very proud to be Iraqi
- Agree that I share a lot in common with Sunni Arabs
- Feel comfortable going to "different" neighborhoods in my town

Index #2: Prospects for Peace

- Believe that it is arbitrary to divide Iraqis in to ethnic and sectarian identities
- Believe that life would be better if Iraqis treated each other as Iraqis first
- Willing to sell land to a non-Christian
- Believe that life these days is unpredictable and dangerous

Index #3: Tolerance

- Believe that most Sunni Arabs disapproved of ISIS
- Describe current friendship group as mixed

Index #4: Muslims as Neighbors

- Would be comfortable with a Sunni Arab as a neighbor
- Would be comfortable with a Sunni Shabak as a neighbor
- Would be comfortable with a Shi'ite Shabak as a neighbor

Index #5: Absolve Muslim Civilians of Blame

- Do not believe that Sunni Arab civilians are responsible for their suffering
- Do not believe that Shabak civilians are responsible for their suffering

[8] These individual items capture: whether Christians should arm themselves for protection, the generalized trust item, an empathy item toward Shabak Muslims (only asked in t1) and an item on preference for playing with mixed vs. homogenous teams in the future (only asked in t1).

4.3 Factor Analysis

I now conduct factor analysis on these seven clusters, to create scores that will serve as t1 outcomes. If fewer than 50% of the variables in a given index for a given observation have missing values, I impute the missing values using medians. If over 50% of the variables in an index have missing values, I drop this observation.

```
# Factor Analysis:
create_factor <- function(data, dv_names, verbose = TRUE) {
  # Keep variables of interest and ensure that they?re numeric
  sub <- data[, dv_names] %>% mutate_all(funs(as.numeric(as.character(.))))
  # Impute missing values with median
  impute <- sub
  impute <- sapply(impute, function(x) ifelse(is.na(x), median(x, na.rm = T), x))
  # Principal component, scaled to have mean 0/sd 1
  f <- princomp(impute, cor = TRUE)
  if (verbose) print(loadings(f))
  dv <- f$scores[, 1]
  dv <- as.numeric(scale(dv))
  # Make sure the variable points the correct way (using question on whether most
  # Muslims approved of ISIS as reference)
if (cor(dv, data$op_isis_t0, use = "complete") < 0) dv <- -1 * dv
# If a row in the original data has more than 50% NAs, then replace the score
# with NA
bool <- apply(sub, 1, function(x) sum(is.na(x)) / ncol(sub) > 0.5)
dv[bool] <- NA
dv }
```

Factor loadings for the seven indices are below. With one exception, every variable loads on to the principal component with a loading of at least 0.42 and up to 0.71.

```
## (1) Coexistence
df$dv_coexist <- create_factor(df, dv_coexist)

Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4
welc_t0       -0.528 -0.441        0.720
empathy_t0    -0.559        -0.778 -0.282
proud_iraqi_t0 -0.567 -0.102 0.601 -0.554
areas_t0      -0.296 0.890  0.156  0.309

              Comp.1 Comp.2 Comp.3 Comp.4
SS loadings     1.00   1.00   1.00   1.00
Proportion Var  0.25   0.25   0.25   0.25
Cumulative Var  0.25   0.50   0.75   1.00

## (2) Prospects for Peace
df$dv_peace <- create_factor(df, dv_peace)
```

```
Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4
secular_t0 -0.579 0.280  0.504  0.577
secular1_t0 -0.396 0.723 -0.334 -0.457
danger_t0   0.538  0.509 -0.333 0.584
land_t0    -0.468 -0.373 -0.724 0.344

              Comp.1 Comp.2 Comp.3 Comp.4
SS loadings    1.00   1.00   1.00   1.00
Proportion Var 0.25   0.25   0.25   0.25
Cumulative Var 0.25   0.50   0.75   1.00


## (3) Tolerance
df$dv_tolerance <- create_factor(df, dv_tolerance)

Loadings:
          Comp.1 Comp.2
friends_t0 0.707 -0.707
op_isis_t0 0.707 0.707

              Comp.1 Comp.2
SS loadings     1.0    1.0
Proportion Var  0.5    0.5
Cumulative Var  0.5    1.0



## (4) Muslims as Neighbors
df$dv_neigh <- create_factor(df, dv_neigh)

Loadings:
               Comp.1 Comp.2 Comp.3
nbr_shabak_sh  -0.420 0.907
nbr_shabak_su  -0.642 -0.297 -0.707
nbr_shabak_su.1 -0.642 -0.297 0.707

              Comp.1 Comp.2 Comp.3
SS loadings    1.000  1.000  1.000
Proportion Var 0.333 0.333 0.333
Cumulative Var 0.333 0.667 1.000

## (5) Absolving Muslims of ISIS Blame

 df$dv_suff <- create_factor(df, dv_suff)

Loadings:
          Comp.1 Comp.2
suff_sh_t0 -0.707 0.707
suff_sa_t0 -0.707 -0.707
```

```
                Comp.1 Comp.2
SS loadings        1.0    1.0
Proportion Var     0.5    0.5
Cumulative Var     0.5    1.0
```

5 Estimation

5.1 Estimating ATE

I use an OLS model with standard errors clustered at the team level and with the covariates mentioned above, as shown below.

```
# Function for estimating treatment effects. Requires package "lfe" for
# clustering the standard errors at the team level.
ate <- function(dta, dv, treat, covars, cluster = "team", extravars_extract = NULL) {

  # Create OLS formula
  rhs <- paste(c(treat, covars), collapse = " + ")
  f <- paste(dv, "~", rhs, "| 0 | 0 |", cluster)
  cat("Formula used:", f, "\n")

  # Estimate regression
  m <- felm(formula(f), data = dta)

  # Extract results of interest
  tidy(m, conf.int = T) %>%
    filter(term %in% c("(Intercept)", treat, extravars_extract))

}
```

The main analyses will be estimated as shown below. I analyze the treatment effect on indices by subtracting the t0 index from the t1 index, and adding demographic controls outlined in section 5.2.

```
# Index #1
ate(df, dv = "dv_coexist", treat = "treat_PAP", covars = covariates)

# Index #2
ate(df, dv = "dv_peace", treat = "treat_PAP", covars = covariates)

# Index #3
ate(df, dv = "dv_ingroup", treat = "treat_PAP", covars = covariates)

# Index #4
ate(df, dv = "dv_neighbors", treat = "treat_PAP", covars = covariates)

# Index #5
ate(df, dv = "dv_safety", treat = "treat_PAP", covars = covariates)
```

```
# Index #6
ate(df, dv = "dv_blame", treat = "treat_PAP", covars = covariates)

# Index #7
ate(df, dv = "dv_trust", treat = "treat_PAP", covars = covariates)
```

Two items only asked in t1: how much one has in common with Shabak Muslims, and whether the respondent prefers to play with co-religionists in future leagues. I re-run the ATE analysis without adjusting for baseline covariates for these two items.

5.2 Covariates to use in Regression Adjustment

I will include the following demographic variables from the baseline pre-survey as covariates to predict the outcome and increase power. These covariates include standard demographic traits, abuse suffered at the hands of ISIS, and whether the player was added or core team member.

```
covariates <- c("age", "city", "kids", "edu", "employment_status",
                "church", "isis_abuse", "added")
```

5.3 Testing Mechanisms

Pettigrew and Tropp (2006) outline three ways in which contact can reduce prejudice. First, contact may increase knowledge of the out-group and "[reveal] negative stereotypes to be false" (Scacco and Warren, 2018, pg. 656). Second, contact may reduce anxiety about interacting with the out-group. Lastly, contact may induce empathy and perspective-taking among participants. I test these mechanisms through corresponding questions in both the t0 and t1 surveys. I test two additional mechanisms: friendship effects (increased tolerance only toward those in the league), and success effects (increased tolerance among those who have the most positive league experience).

- **Filling information gaps about the "other."** Contact may reduce prejudice by providing first-hand information about the out-group that overturn stereotypes (Peffley, Hurwitz and Sniderman, 1997). If contact increases information about Muslims, then treated Christians should be more likely than control Christians to reject stereotypical statements like "most Muslims approved of ISIS." We would also see amplified effects among those with reporting the least contact with Muslims t0, described in Section 5.4.

- **Increased empathy.** By humanizing out-groups, contact allows participants to put themselves in the shoes of the "other" and see similarities rather than differences (Simonivitz, Kezdi and Kardos, 2017). If the increased empathy mechanism is true, we would see a treatment effect on the empathy item.

- **Reduced out-group anxiety.** Inter-group contact may reduce anxiety regarding out-group encounters. I measure this using a survey item on "sometimes" or "often" feeling comfortable visiting out-group neighborhoods, and patronage of restaurants in out-group areas.

- **Friendship effects.** Social psychologists caution that contact may "individualize" those one meets rather than lead to generalized tolerance. One begins to see a particular out-group

member as an individual rather than associate them with the out-group at all. If this is the case, then the intervention would improve outcomes only toward Muslims encountered in the league but not in general. Under the friendship effect mechanism, we would see movement on the event attendance and top sportsman vote outcomes, but not on the remainder of outcomes aimed capturing attitudes toward Muslims more broadly.

- **Success and enjoyment.** Allport's theory stipulates that inter-group contact must be positive in order to build tolerance. I measure a positive experience with Muslims three ways: (1) treated teams that advance to the quarter-final knock-out stage, (2) treated teams that received extra players in the top 25^{th} skill percentile, as coded by research staff, and (3) players on treated teams that say they would like to participate in the league again in the t1 survey.[9]

5.4 Treatment Effect Heterogeneity by Subject Attributes

Detecting interaction effects typically requires a ten-fold increase (or more) in the sample size needed to estimate a main effect. Interaction effects are therefore only suggestive. I expect that positive ATE estimates would be amplified among three groups:

1. Those who report a friendship group generally consisting of those "who belong to the same group as me" at t0.
2. Those with the most positive experience with Muslims in the league, defined as reaching the quarterfinals, having added players in the top 25^{th} skill percentile, and indicating that they would like to participate in another league in the future.
3. Those in the 25^{th} percentile of disagreeing or strongly disagreeing that they have "a lot in common with Sunni Arabs" at t0 (randomization was blocked on this variable).

As an example, the code below demonstrates how I will estimate heterogenous effects for those with previous Muslim friends and those on successful teams:

```
# Heterogeneous effects by having a homogenous friendship group
covariates_new <- c(covariates, "friends_mixed")

# Interacting previous contact with the treatment effect, main dv analysis
ate(df, dv = "dv_trust", treat = "treat_PAP", covars = covariates_new,
    extravars_extract = c("friends_mixed"))

 # Heterogeneous effects by t0 empathy toward Muslims (
covariates_new <- c(covariates, "empathy_t0")

# Interacting previous contact with the treatment effect, main dv analysis
ate(df, dv = "dv_trust", treat = "treat_PAP", covars = covariates_new,
    extravars_extract = c("empathy_t0"))

# Heterogeneous effects by team success, skill level of added players. and league
    experience, main dv analysis
```

[9]This final sub-group is merely suggestive, however, as it relies on conditioning on post-treatment variables.

```
covariates_new <- c(covariates, "quarter_finals", "skilled_extra_players", "
    enjoyed_league")

# Interacting heterogenous trait with the treatment effect
ate(df, dv = "dv_trust", treat = "treat_PAP", covars = covariates_new,
    extravars_extract = c("quarter_finals", "skilled_extra_players", "enjoyed_league"))
```

6 Other

6.1 Enumerator Heterogeneity

The experimental league surveys were self-administered on Qualtrics using tablets. The 117 comparison league surveys were administered in hard copy format under research assistant supervision. Research assistants then input the responses into Qualtrics, a random sample of which was cross-validated to confirm cross-enumerator reliability. In robustness tests, I will control for enumerator fixed effects and conduct additional tests of whether the ATE varies by enumerator on average. There is no ex-ante reason to assume enumerator heterogeneity, however, as the surveys to be input were randomly distributed across research assistants.

6.2 Non-Compliance and Attrition

All participants in Wave 1 of the experiment complied with protocol, largely because of the semi-professional nature of the existing teams. Non-compliance may occur in Wave 2, however. For this, I make the no-defiers assumption. I will adjust the point estimates above to account for non-compliance by dividing the point estimates by the proportion of treated compliers in the treatment group minus the proportion of treated compliers in the contacted control group.

Wave 1 yielded a retention rate of 86.4% of contacted participants. The remaining 14.6% initially agreed to participate, but were forced to drop out early on (before treatment assignments were made) due to housing constraints. Interviews with those who dropped out reveal that the Baghdad government required displaced people whose homes had been sufficiently reconstructed to return in order to maintain access to public service, and so some players were forced to return to their hometowns. There is thus a convincing case for random missingness, as attrition cannot be predicted by pre-treatment characteristics.

6.3 Missing Values

Missing values are largely limited to the paper surveys administered in the comparison league, as the tablet surveys requested responses for all questions. I impute missing baseline values using the default predictive mean matching method in the `mice` package. I will present the results with and without this imputation as a robustness check. All variables are taken as predictors for the imputation algorithm, except player ID.

6.4 Pooled Pilot Results

Because the intervention replicates the pilot fielded in 2017, I will present pooled results to increase the sample size, as well as the scaled-up RCT results separately. The indices will be re-created

using the pilot data to the closest degree possible for this analysis.

6.5 Power Calculation

Pilot results have allowed me to conduct design effect analyses with the following conservative specifications: 80% power, 95% C.I., an effect size of 10 percentage-points with a mean portion of 0.5 in the control group (most outcomes, like attending a mixed social event, are binary), and 14 players per cluster. After calculating *rho* for the primary outcomes, the most conservative estimate for the minimum effective sample size given the team-clustered design is 376 ? well within the proposed sample size of 444 participants (in addition to at least 300 household members and 37 coaches).

Below is an example of how the power calculation was conducted using the pilot data:

```
library(ImPerm)

## power calculator

pwr.2p.test(h = ES.h(p1 = 0.50, p2 = 0.6), sig.level = 0.05, power = .80)

     Difference of proportion power calculation for binomial distribution (arcsine
        transformation)

              h = 0.2013579
              n = 387.1677
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes

## calculating design effect while adjusting for clusters (where chr$dv_coexist is the
    coexistence index outcome among christians)

> deff(chr$dv_coexist, chr$team.x)
          n     clusters        rho        deff
168.00000000 14.00000000 0.01381732  1.15659630

## formula: (1 + rho *(14 -1))

> 1 + (0.01381732*13)
[1] 1.179625

## effective sample size

> 444/ 1.179625
[1] 376.3908
```

6.6 Match Infractions and Referee Identity

The assignment of referee identity is random, leading to a natural experiment. I analyze match-level data on yellow cards and red cards to test whether referee religious identity shapes aggression.

References

Enos, Ryan D. 2017. *The space between us: Social geography and politics*. Cambridge University Press.

Lowe, Matt. 2017. "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration." *MIT Working Paper* .

Peffley, Mark, Jon Hurwitz and Paul M Sniderman. 1997. "Racial stereotypes and whites' political views of blacks in the context of welfare and crime." *American Journal of Political Science* pp. 30–60.

Pettigrew, Thomas F and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751.

Scacco, Alexandra and Shana S Warren. 2018. "Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria." *American Political Science Review* pp. 1–24.

Simonivitz, Gabor, Gabor Kezdi and Peter Kardos. 2017. "Seeing the World Through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice." *American Political Science Review* pp. 1–8.