LA County Voter File to Email Panel: Pre-Analysis Plan *

Bryan Wilcox-Archuleta *University of California, Los Angeles*

Background and Overview

Public opinion researchers in the U.S. often face a number of limitations in studying the opinions of the mass public. For one, the overall response rates of telephone based interviews has been declining. Second, the cost of telephone based interviews continually increases. Third, cross-sectional surveys are less and less suitable to study many of the questions that researchers care about. Fourth, despite the development and proliferation of on-line panels, the representativeness of these platforms remains worrisome, especially if scholars are interested in non-general population based studies (Berinsky, Huber and Lenz, 2012, Coppock (2017)). Fifth, despite the proliferation of on-line panels, these methods still remain out of reach for many researcher given their non-trivial costs.

To overcome many of these limitations in public opinion research, I test the feasibility of voter file based email-to-panel recruiting (VETP). VETP offers many advantages over existing practices. For one, voter files with email addresses are becoming more available. Second, voter files are relatively low cost, allowing access to survey tools to a broad range of researchers, including graduate students and researchers outside of research universities. Third, unlike online panels, it is possible to conduct true probability samples from voter files. Voter files contain the universe of registered voters in a jurisdiction, allowing for a researcher to sample from that population with a known probability. While not all voters provide email addresses, calculating the probability of inclusion is still possible. Because the population parameters are known, it is also possible to weight the response back to a know populations, allowing population based inferences, something the is by definition impossible in online panel based samples. While online panels have been shown to perform well in terms of representativeness Vavreck and Rivers (2008), we still know little about the underlying population or much about those who agree and subsequently participate in these panels. Recent discussions of survey bots and semi-professional survey takers continue to cast

*I would like to thank Erin Hartman, Tyler Reny, Joy Wilke, Loren Collingwood, Matt Barreto for helpful insight and feedback

doubt on the quality of these methods.

Given these benefits, the goal of this study is to asses the cost, feasibility, and overall use of VETP. The remainder of this PAP outlines the research questions, quantities of interest, and displays the code used to determine these quantities. One of the key questions here is how well VETP methods work for non-white respondents. To accomplish this, I have developed a sampling strategy that increases the probability that non-whites are included in the sample. I also conduct this sample in Los Angeles County, CA, offering vast diversity in terms of racial and ethnic composition.

Research Questions

- What is the response rate for VETP?
- What are the response rate by racial/ethnic groups for VETP?
- What are the demographic characteristics of respondents who opt in to be in a panel?
- Are those who opt in different from the remaining population in terms of demographic characteristic?
- What are the attrition rates for panel members?
- Do a number of studies replicate in the VETP sample?
- Do these studies replicate in those who agree to participate in a panel?

Quantities of Interest

The research questions above have identified key quantities of interest that I am interested in estimating. In this section I discuss each quantity and outline how I plan to estimate them.

Overall Response Rate

The overall response rate will be calculated following AAPOR's guidelines. Each respondent will be sent an initial email then sent the first follow up email three days later. The final follow up email will be sent 5 days after the second follow up email, 8 days after the original email. I will also calculate the response rates of each blast. Not all emails are valid and as such, I have devised a system of dealing with different responses.

- N=Total Sample Used
- I=Complete Interviews - I are respondents that went through the full survey as well as those who answered all the questions but may have closed out of the last question before clicking next. The last question thanks them for their time, provides a debrief, and takes comments. If they don't click next here, the interview is not recorded as complete in Qualitrics.
- P=Partial Interviews - P are respondents who answer at least 1 question.
- R=Refusal and break off - R are respondents who open the link from the email text, but do not answer a question.
- NC=Non Contact - NC are respondents where an email was delivered but they did not respond.
- O=Other
- UH=Unknown Household
- UO=Unknown Other - UO are responses that are generally returned. For this project this will include bounces, returned emails with forwarding information.

```
library(tidyverse)
library(responserates)
# overall
rate_list <- list(i = i, p = p, r = r, nc = nc, o = o, uh = uh,
    uo = uo)
rates(rate_list)


# blast 1
rate_list_1 <- list(i = i, p = p, r = r, nc = nc, o = o, uh = uh,
    uo = uo)
rates(rate_list_1)


# blast 2
rate_list_2 <- list(i = i, p = p, r = r, nc = nc, o = o, uh = uh,
    uo = uo)
```

```
rates(rate_list_2)


# blast 3
rate_list_3 <- list(i = i, p = p, r = r, nc = nc, o = o, uh = uh,
    uo = uo)
rates(rate_list_2)
```

Response Rate by Race/Ethnic Group

I will test the response rate by each of the racial and ethnic groups.

```
# white
rate_list_white <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_white)


# black
rate_list_black <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_black)


# latino
rate_list_latino <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_latino)


# aapi
rate_list_aapi <- list(i = i, p = p, r = r, nc = nc, o = o, uh = uh,
    uo = uo)
rates(rate_list_aapi)
```

```r
# other
rate_list_other <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_other)
```

Panel Agreement Rate

This will be calculated as the proportion of completed survey who agree to participate in future studies.

```r
mean(df$panel_recruit, na.rm = T)
```

Wave 2 Response Rate

To measure panel attrition, I will re-contact all those who opted into the panel 60 days later. I will use the AAPOR guidelines. Two follow up emails will be sent within 10 days after the initial email.

```r
# wave 2
rate_list_wave_2 <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_wave_2)
```

Wave 3 Response Rate

To measure panel attrition, I will re-contact all those who opted into the panel 120 days later. I will use the AAPOR guidelines. Two follow up emails will be sent within 10 days after the initial email.

```r
# wave 3
rate_list_wave_3 <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_wave_3)
```

Wave 4 Response Rate

To measure panel attrition, I will re-contact all those who opted into the panel 180 days later. I will use the AAPOR guidelines. Two follow up emails will be sent within 10 days after the initial email.

```
# wave 4
rate_list_wave_4 <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_wave_4)
```

Wave 5 Response Rate

To measure panel attrition, I will re-contact all those who opted into the panel 365 days later. I will use the AAPOR guidelines. Two follow up emails will be sent within 10 days after the initial email.

```
# wave 5
rate_list_wave_5 <- list(i = i, p = p, r = r, nc = nc, o = o,
    uh = uh, uo = uo)
rates(rate_list_wave_5)
```

Replication of Kahneman and Tversky (1979)

Kahneman and Tversky (1979) provide the main replication exercise for this project. The goal is to replicate the broad findings presented in the 1979 paper in the LA county voter population. This study has been replicated in a wide variety of contexts and I expect to see the results in LA county to follow. Below is the code used to test the replication from (Kahneman and Tversky, 1979).

```
# asian disease problem
prop.table(table(df$asian_disease_1))
prop.test(prop.table(table(df$asian_disease_1)), n = table(df$asian_disease_1),
    correct = F)
```

```r
prop.table(table(df$asian_disease_2))
prop.test(prop.table(table(df$Q30)), n = table(df$Q30), correct = F)

# Problem 1 and Problem 2
prop.table(table(df$problem_1))
prop.test(prop.table(table(df$problem_1)), n = table(df$problem_1),
    correct = F)
binom.test(table(df$problem_1))

prop.table(table(df$problem_2))
prop.test(prop.table(table(df$problem_2)), n = table(df$problem_2),
    correct = F)
binom.test(table(df$problem_2))

# Problem 3 and Problem 4
prop.table(table(df$problem_3))
prop.test(prop.table(table(df$problem_3)), n = table(problem_3),
    correct = F)
binom.test(table(df$problem_3))

prop.table(table(df$problem_4))
prop.test(prop.table(table(df$problem_4)), n = table(df$problem_4),
    correct = F)
binom.test(table(df$problem_4))

# Problem 5 and Problem 6
prop.table(table(df$problem_5))
prop.test(prop.table(table(df$problem_5)), n = table(df$problem_5),
    correct = F)
```

```r
binom.test(table(problem_5))


prop.table(table(df$problem_6))
prop.test(prop.table(table(df$problem_6)), n = table(df$problem_6),
    correct = F)
binom.test(table(df$problem_6))


# Problem 7 and Problem 8
prop.table(table(df$problem_7))
prop.test(prop.table(table(df$problem_7)), n = table(df$problem_7),
    correct = F)
binom.test(table(problem_7))


prop.table(table(df$problem_7))
prop.test(prop.table(table(df$problem_7)), n = table(df$problem_7),
    correct = F)
binom.test(table(df$problem_7))


# problem 9
prop.table(table(df$problem_9))
prop.test(prop.table(table(df$problem_9)), n = table(df$problem_9),
    correct = F)
binom.test(table(df$problem_9))


# problem 13 and problem 13'
prop.table(table(df$problem_13))
prop.test(prop.table(table(df$problem_13)), n = table(df$problem_13),
    correct = F)
binom.test(table(df$problem_13))
```

```r
prop.table(table(df$problem_13_prime))
prop.test(prop.table(table(df$problem_13_prime)), n = table(df$problem_13_prime),
    correct = F)
binom.test(table(df$problem_13_prime))

# problem 14 and problem 14'
prop.table(table(df$problem_14))
prop.test(prop.table(table(df$problem_14)), n = table(df$problem_14),
    correct = F)
binom.test(table(df$problem_14))

prop.table(table(df$problem_14_prime))
prop.test(prop.table(table(df$problem_14_prime)), n = table(df$problem_14_prime),
    correct = F)
binom.test(table(df$problem_14_prime))
```

Sampling

I use a stratified sampling procedure with disproportionate allocation to ensure large enough samples of each strata. I use four stratification variables:

1. Age
    - 18-29
    - 30-44
    - 45-64
    - 65+
2. Predicted Race
    - White

- Black
- Latino
- AAPI
- Other

3. Party ID

   - Democrat
   - Republican
   - Other

4. Vote in 2016 General

   - Voted
   - Did not vote

This results in 120 unique strata. Based on pilot work and conversations with researchers, I over sample younger and non-white voters. I draw a sample of 130,000 respondents. Below are the predicted cell sizes from the sample code show below:

```
df %>% group_by(age_bucket, race_bucket, party_id_3, voted_2016_general) %>%
    summarise(strata_size = n()) %>% kable(., caption = "Estimated Strata Cell Sizes")
```

Table 1: Estimated Strata Cell Sizes

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
1	1	democrat	0	3157
1	1	democrat	1	3228
1	1	other	0	3045
1	1	other	1	3100
1	1	republican	0	1505
1	1	republican	1	1565
1	2	democrat	0	2601

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
1	2	democrat	1	2746
1	2	other	0	2207
1	2	other	1	2051
1	2	republican	0	539
1	2	republican	1	421
1	3	democrat	0	615
1	3	democrat	1	619
1	3	other	0	671
1	3	other	1	691
1	3	republican	0	618
1	3	republican	1	669
1	4	democrat	0	2583
1	4	democrat	1	2844
1	4	other	0	2600
1	4	other	1	2643
1	4	republican	0	1267
1	4	republican	1	1316
1	5	democrat	0	335
1	5	democrat	1	336
1	5	other	0	314
1	5	other	1	317
1	5	republican	0	326
1	5	republican	1	332
2	1	democrat	0	2968
2	1	democrat	1	3270
2	1	other	0	2803
2	1	other	1	2998
2	1	republican	0	1364

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
2	1	republican	1	1609
2	2	democrat	0	2355
2	2	democrat	1	2869
2	2	other	0	1919
2	2	other	1	2354
2	2	republican	0	478
2	2	republican	1	655
2	3	democrat	0	681
2	3	democrat	1	708
2	3	other	0	635
2	3	other	1	646
2	3	republican	0	658
2	3	republican	1	664
2	4	democrat	0	2566
2	4	democrat	1	2972
2	4	other	0	2687
2	4	other	1	2884
2	4	republican	0	1173
2	4	republican	1	1473
2	5	democrat	0	332
2	5	democrat	1	333
2	5	other	0	341
2	5	other	1	368
2	5	republican	0	343
2	5	republican	1	337
3	1	democrat	0	1633
3	1	democrat	1	1948
3	1	other	0	850

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
3	1	other	1	960
3	1	republican	0	709
3	1	republican	1	845
3	2	democrat	0	1550
3	2	democrat	1	1827
3	2	other	0	752
3	2	other	1	875
3	2	republican	0	407
3	2	republican	1	655
3	3	democrat	0	328
3	3	democrat	1	303
3	3	other	0	322
3	3	other	1	350
3	3	republican	0	333
3	3	republican	1	324
3	4	democrat	0	1484
3	4	democrat	1	1791
3	4	other	0	873
3	4	other	1	935
3	4	republican	0	757
3	4	republican	1	880
3	5	democrat	0	305
3	5	democrat	1	346
3	5	other	0	322
3	5	other	1	303
3	5	republican	0	332
3	5	republican	1	308
4	1	democrat	0	548

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
4	1	democrat	1	837
4	1	other	0	352
4	1	other	1	616
4	1	republican	0	210
4	1	republican	1	589
4	2	democrat	0	579
4	2	democrat	1	873
4	2	other	0	284
4	2	other	1	548
4	2	republican	0	102
4	2	republican	1	370
4	3	democrat	0	636
4	3	democrat	1	663
4	3	other	0	614
4	3	other	1	639
4	3	republican	0	624
4	3	republican	1	637
4	4	democrat	0	590
4	4	democrat	1	835
4	4	other	0	658
4	4	other	1	816
4	4	republican	0	464
4	4	republican	1	767
4	5	democrat	0	301
4	5	democrat	1	355
4	5	other	0	301
4	5	other	1	323
4	5	republican	0	249

age_bucket	race_bucket	party_id_3	voted_2016_general	strata_size
4	5	republican	1	309

Generalizibility

Those who provide their email address on the voter file are likely different on both observed characteristics and unobserved characteristics. Further, those who respond to a VETP recruitment could be different than those who do not. The overarching goal is to make population based inferences. In this PAP, I propose to weight respondents back to two populations: 1) those with active emails on the file and 2) the entire registered voter population. I plan to weight on four key variables: Age of respondent, predicted race, party id, and whether respondents voted in the general election.

```
voter <- voter %>% dplyr::select(voter_id, clean_email, age_bucket,
    race_bucket, voted_2016_general, party_id_3) %>% na.omit()

# survey design population object for creating targets
pop_design <- svydesign(~1, data = voter)

# completed sample
sample <- df %>% select(email, race_bucket, age_bucket, voted_2016_general,
    party_id_3) %>% na.omit()



# survey object for the sample
survey_design <- svydesign(~1, data = sample, weights = sample$pik)




# vars we are weighting on
```

```
weight_vars = c("race_bucket", "age_bucket", "voted_2016_general",
    "party_id_3")



### function for generating population targets
gen_xtabs = function(vars, data, margins = TRUE, percent = TRUE) {

    if (!margins) {
        if (class(data)[1] == "survey.design2") {
            tabs = as.data.frame(svytable(as.formula(paste0("~",
                paste(vars, collapse = " + "), " - 1")), data))
        } else {
            warning("Sample object must be a survey.design object.")
            break
        }
        if (percent)
            tabs$Freq = tabs$Freq/sum(tabs$Freq)
    } else {
        if (class(data)[1] == "survey.design2") {
            tabs = NULL
            for (var in vars) {
                summary = as.data.frame(svymean(as.formula(paste0("~",
                  var, " - 1")), data))
                temp_tabs = as.data.frame(t(summary$mean))
                names(temp_tabs) = rownames(summary)
                if (var != vars[1] & length(vars) > 1) {
                  temp_tabs = temp_tabs[-1]
                }
                tabs = c(tabs, temp_tabs)
```

```
            }
            tabs = as.data.frame(tabs)
        } else {
            warning("Population object must be a survey.design object.")
            break
        }
        if (!percent)
            tabs = tabs * nrow(data)
    }

    if (margins) {
        tabs = t(tabs)
    }
    return(tabs)
}

combine_tabs = function(vars, pop, samp, weight = NULL, margins = TRUE) {
    samp_tabs = gen_xtabs(vars, samp, margins)
    pop_tabs = gen_xtabs(vars, pop, margins)
    if (!is.null(weight)) {
        weight_tabs = gen_xtabs(vars, weight, margins)
    }
    if ("Freq" %in% names(pop_tabs)) {
        names(samp_tabs)[ncol(samp_tabs)] = "Samp_Prop"
        samp_tabs$Pop_Prop = pop_tabs$Freq
        if (!is.null(weight)) {
            samp_tabs$Weight_Prop = round(weight_tabs$Freq, 4)
        }
    } else {
```

```
        samp_tabs = data.frame(Var = rownames(samp_tabs), Samp_Prop = samp_tabs,
            Pop_Prop = pop_tabs)
        if (!is.null(weight)) {
            samp_tabs$Weight_Prop = round(weight_tabs, 4)
        }
    }

    samp_tabs$Samp_Prop = round(samp_tabs$Samp_Prop, 5)
    samp_tabs$Pop_Prop = round(samp_tabs$Pop_Prop, 5)

    samp_tabs$Diff = round(samp_tabs$Pop_Prop - samp_tabs$Samp_Prop,
        5)
    if (!is.null(weight)) {
        samp_tabs$Weight_Prop = round(samp_tabs$Weight_Prop,
            5)
        samp_tabs$Diff_Weight = round(samp_tabs$Pop_Prop - samp_tabs$Weight_Prop,
            5)
    }
    rownames(samp_tabs) = NULL
    return(samp_tabs)
}

# rake
survey_rake = calibrate(survey_design, as.formula(paste0("~",
    paste(weight_vars, collapse = " + "), " - 1")), gen_xtabs(weight_vars,
    pop_design, margins = T, percent = TRUE), calfun = "raking",
    maxit = 1000, epsilon = 1e-04)

bal_table <- as.data.frame(combine_tabs(weight_vars, pop_design,
```

```
    survey_design, survey_rake))

write_csv(bal_table, "manuscript/tables/calibration_table.csv")  # balance table

# get the weights
sample$wts <- weights(survey_rake)

# merge weights back to our sample
res_wtd <- left_join(df, sample, by = "email") %>% mutate(voter_id = as.character(voter_id))



# add census tracts from voter file
census_tract <- voter %>% dplyr::select(voter_id, census_tract,
    voted_primary, voted_general) %>% mutate(voter_id = as.character(voter_id))

# write final data set
res_final <- left_join(res_wtd, census_tract, by = "voter_id") %>%
    write_csv("data/clean_data_weighted.csv")
```

Publications

There are two planned publications based off this work. The first is a write up of the results will report the findings from the replication and the first wave. This paper will show the feasibility and generalizability of this method along with recontact rates and panel attrition. The second paper, nearly a year later, will show the feasibility of this method in terms of developing on online panel from registered voters.

References

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's mechanical turk." *Political Analysis* 20(3):351–368.

Coppock, Alexander. 2017. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* .

Kahneman, D and A Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica: Journal of the Econometric Society* 47(2):263–292.
URL: *http://www.jstor.org/stable/10.2307/1914185*

Vavreck, Lynn and Douglas Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion & Parties* 18(4):355–366.