Comment on White, Nathan, and Faller (2015)

Alexander Coppock, Yale University

June 19, 2017

White, Nathan, and Faller (2015) present the results of a (tremendous, very well done, and important!) audit experiment in which county election officials are sent emails from putatively White or Latino names asking "I've been hearing a lot about voter ID laws on the news. What do I need to do to vote?" Whereas White names received a response 70.5% of the time, Latino names were responded to 64.8% of the time, for a large and statistically significant average treatment effect estimate of negative 5.7 percentage points.

In an additional analysis, the authors examine whether the ethnicity of the name changed the tone (the accuracy and friendliness) of the emails that were sent in response. The authors estimate the treatment effects on tone conditioning on response. This procedure is prone to bias because response is a post-treatment outcome (Aronow, Baron, and Pinson, in press; Montgomery, Nyhan, and Torres, 2016). As noted by the original authors in their footnote 29, the substantive results of the experiment do not depend on this choice – the purpose of this comment is emphatically not to overturn or in any way challenge the conclusions of White, Nathan, and Faller (2015), but rather to take it as an instructive example.

In this setting, a subject might be one of the four types in the table below. $R_i(Z)$ is the response potential outcome depending on whether subject i is assigned to a non-Latino name ($Z = 0$) or a Latino name ($Z = 1$). A types always respond, D types never respond, and B and C types respond if and only if they are in one condition or the other. $Y_i(Z)$ is the tone potential outcome and is **undefined** if a subject does not respond. We would like to estimate the effect of the name on the tone for A types only because the effect among other types is undefined. Our most-interesting estimand is therefore $\mathrm{E}\left[Y_i(1) - Y_i(0) | R_i(0) = R_i(1) = 1\right]$. Unfortunately, the design will not reveal enough information to estimate this estimand because we don't know who the A types are. In the Non-Latino name group, responders are As and Cs, whereas in the Latino name group, responders are As and Bs.

Table 1: Types of Subjects

Type	$R_i(0)$	$R_i(1)$	$Y_i(0)$	$Y_i(1)$	$Y_i^*(0)$	$Y_i^*(1)$
A	1	1	$Y_i(0)$	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$
B	0	1	NA	$Y_i(1)$	0	$Y_i(1)$
C	1	0	$Y_i(0)$	NA	$Y_i(0)$	0
D	0	0	NA	NA	0	0

A way around this is to redefine the outcome variable to be $Y_i^*(Z)$, which is equal to $Y_i(Z)$ if $R_i^*(Z) = 1$ and 0 otherwise. Crucially, this means that emails never sent are "not friendly." They are also "not ambiguous" and "not noninformative." We can then estimate a coherent estimand, the average effect of treatment on this new dependent variable: $\mathrm{E}\left[Y_i^*(1) - Y_i^*(0)\right]$.

Conditioning on response biases our estimates of either estimand because the treatment groups are no longer equivalent. However, under standard experimental assumptions, we can generate unbiased estimates of the second estimand.

White, Nathan, and Faller (2015) report the results of their analysis on page 136. Table 2 reproduces the difference-in-means estimate for all tone dependent variables (both the "accuracy" DV, which can take on 6 values that I am analyzing as 6 separate binary variables and "friendliness", which is also binary.) If we do condition on response, we find that Latino names obtain 4.5 percentage points fewer accurate emails, 4.2 percentage points more non-informative emails and 3.5 percentage points fewer friendly emails.

Table 3 shows the same analysis without conditioning on response. The effect on accuracy is now 6 percentage points, the effect on non-informative emails descends to 1.5 percentage points and is no longer significant, while the effect on the friendliness nearly doubles to negative 6.1 percentage points. As it happens, the sign and significance of the estimates in this experiment do not change much when we do not condition on a post-treatment outcome, but that was just good luck. Conditioning on post-treatment outcomes can lead to bias of unpredictable sign and magnitudes and should be avoided.

Table 2: Effects on tone that do condition on response.

	abs_acc	ambiguous	contains_inacc	general	narrow	non_informative	friendly_no_na
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
first_name_latino	-0.045^{**}	-0.003	0.004	0.008	-0.005	0.042^{**}	-0.035^{*}
	(0.021)	(0.008)	(0.004)	(0.013)	(0.011)	(0.018)	(0.020)
Constant	0.550	0.040	0.009	0.105	0.075	0.220	0.677
	(0.015)	(0.006)	(0.003)	(0.009)	(0.008)	(0.012)	(0.014)
N	2,162	2,162	2,162	2,162	2,162	2,162	2,162
R^2	0.002	0.0001	0.0003	0.0002	0.0001	0.002	0.001

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$
HC2 Robust standard errors are in parentheses.

Table 3: Effects on tone that do not condition on response.

	abs_acc	ambiguous	contains_inacc	general	narrow	non_informative	friendly_no_na
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
first_name_latino	-0.060^{***}	-0.004	0.002	-0.001	-0.008	0.015	-0.061^{***}
	(0.017)	(0.006)	(0.003)	(0.009)	(0.008)	(0.013)	(0.018)
Constant	0.388	0.028	0.006	0.074	0.053	0.155	0.477
	(0.012)	(0.004)	(0.002)	(0.007)	(0.006)	(0.009)	(0.013)
N	3,195	3,195	3,195	3,195	3,195	3,195	3,195
R^2	0.004	0.0002	0.0001	0.00000	0.0003	0.0004	0.004

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$
HC2 Robust standard errors are in parentheses.

References

Aronow, P. M., J. Baron, and L. Pinson (In Press). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*.

Montgomery, J. M., B. Nyhan, and M. Torres (2016). How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do About It. *Unpublished Manuscript*.

White, A. R., N. L. Nathan, and J. K. Faller (2015). What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials. *American Political Science Review 109*(1), 129–142.

```
library(tidyverse)
library(commarobust)
library(stargazer)
load("voterIDexp_data_sept2014.Rdata")

# Subset to VR emails in week one, knock out two states
subset_for_use <- data %>%
  filter(!code %in% c(41, 12), week == 1, first_text == 1)

# Recode so no NAs
subset_for_use <-
  within(subset_for_use,{
    bounced <- as.numeric(is.na(response))
    response_no_na <- response
    response_no_na[is.na(response)] <- 0
    accuracy_no_na <- as.character(accuracy)
    accuracy_no_na[accuracy == ""] <- "No Response"
    abs_acc <- as.numeric(accuracy_no_na == "Absolutely accurate")
    ambiguous <- as.numeric(accuracy_no_na == "Ambiguous")
    contains_inacc <- as.numeric(accuracy_no_na == "Contains inaccurate information")
    general <- as.numeric(accuracy_no_na == "General, but accurate")
    narrow <- as.numeric(accuracy_no_na == "Narrow, but accurate")
    non_informative <- as.numeric(accuracy_no_na == "Non-informative response")
    friendly_no_na <- friendly
    friendly_no_na[is.na(friendly)] <- 0
  })

fit_1 <- lm(abs_acc ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_2 <- lm(ambiguous ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_3 <- lm(contains_inacc ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_4 <- lm(general ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_5 <- lm(narrow ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_6 <- lm(non_informative ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))
fit_7 <- lm(friendly_no_na ~ first_name_latino, data = filter(subset_for_use, response == 1 & bounced == 0))

stargazer(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7,
          se = makerobustseslist(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7),
          p = makerobustpslist(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7),
          style = "apsr",
          omit.stat = c("f", "ser", "adj.rsq"),
          notes = c("HC2 Robust standard errors are in parentheses."),
          title = "Effects on tone that do condition on response.",
          label = "tab:conditional",
          out = "conditional.tex")

fit_1 <- lm(abs_acc ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_2 <- lm(ambiguous ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_3 <- lm(contains_inacc ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_4 <- lm(general ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_5 <- lm(narrow ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_6 <- lm(non_informative ~ first_name_latino, data = filter(subset_for_use, bounced == 0))
fit_7 <- lm(friendly_no_na ~ first_name_latino, data = filter(subset_for_use, bounced == 0))

stargazer(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7,
          se = makerobustseslist(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7),
          p = makerobustpslist(fit_1, fit_2, fit_3, fit_4, fit_5, fit_6, fit_7),
          style = "apsr",
          omit.stat = c("f", "ser", "adj.rsq"),
          notes = c("HC2 Robust standard errors are in parentheses."),
          title = "Effects on tone that do not condition on response.",
          label = "tab:unconditional",
          out = "unconditional.tex")
```