COOPERATIVE
ELECTION
STUDY

Do Repeat CES Respondents Affect Inferences?
A Preliminary Report

Brian Schaffner, co-PI
Cooperative Election Study
Tufts University

9/26/2022

Introduction

In this memo, I explore the implications of repeat respondents for making inferences with the Cooperative Election Study (CES) data. As noted in a recent report by Deshpande and Cha (2020), approximately 25% of the 2020 CES respondents had also taken the 2018 CES survey. That memo correctly warns that combining multiple years of the CES will mean including a significant number of repeat respondents which means researchers should largely treat the CES as separate cross-sections (rather than as a single cumulative sample) in their analyses.

But does the presence of repeat respondents matter for inferences drawn from the CES data in a given year? Repeat respondents are common for online surveys that rely on panels and because of the size of the CES it is especially necessary to use some respondents who have taken a prior CES survey. The Pew Research Center, which maintains its own panel (the American Trends Panel), found little evidence in a 2021 report that repeated survey taking affected responses or behavior among their panelists. Here I consider whether repeat respondents are likely to cause any inferential problems for researchers using the CES using the 2020 survey as a case study.

Identifying likely repeaters

YouGov cannot release the caseids for repeat respondents due to concerns about personally identifiable information (PII). However, to produce a list of likely repeaters, I matched 2018 CES respondents to the 2020 file on zip code, birth year, and gender. I use these three variables because (1) they are mostly likely to be stable across a two-year period and (2) they are reasonably unique to survey respondents. To be sure, zip code, birth year, and gender are not unique to most individuals in the population (Ansolabehere and Hersh 2017), and there are even some respondents in single CES cross-sections who share values on those three variables. But in the 2020 CES, about 94% of respondents had a unique zip code, birth year, and gender combination.

This method produced a list of 16,508 individuals who took the 2020 CES who had an identical zip code, birth year, and gender as a 2018 respondent. I tag these individuals as "likely repeaters." Note that YouGov indicated that at least 15,000 2020 CES respondents had also taken the 2018 survey, so the number of 2020 respondents who match on birth year, zip code, and gender is a bit higher than the YouGov figure. In other words, it is likely that this group includes some false positives (respondents who did not take the survey in 2018), particularly given that having a shared birth year, zip code, and gender did not even uniquely identify every respondent within a given year.

To provide a benchmark for how widespread the false positive rate might be, I use the 2010-2012 CCES panel survey, which includes 18,000 respondents who were interviewed in both 2010 and 2012. These 18,000 respondents are known to be the same individuals, so we can use their stability on responses to other questions as a guide. Specifically, I look at the percentage who gave the same response to three demographic items: (1) race/ethnicity, (2) marital status, and (3) whether they had children under the age of 18. The table below compares the rate at which "likely repeaters" gave the same responses to these items in 2018 and 2020 compared to the benchmark stability rate for the 2010-2012 panel survey.

Percent giving same response to…	Likely repeaters	Benchmark 2010-12 panel
Race	90.0%	95.3%
Marital status	84.6%	92.6%
Kids under 18	91.5%	94.1%

The table shows that likely repeaters are somewhat less consistent in their survey responses compared to known repeat respondents from the 2010-2012 panel survey. However, the discrepancies are not especially large and suggest that while there are some false positives among the group (probably about 5% to 7%), the vast majority of individuals identified are indeed likely to be respondents who also took the 2018 survey.

There are also likely to be a few false negatives with this approach. In particular, if a repeat respondent moved and changed zip codes between 2018 and 2020 they would not be identified as a "likely repeater." About 10% of individuals change addresses within a two-year period, though not all of these moves would produce a change in zip code. If moving is unrelated to being a repeat respondent, then this would create approximately 1,500 false negatives. However, it is likely that a disruption such as moving would also make an individual more likely to drop off of the YouGov panel (and therefore less likely to be a repeat respondent), which means the false negative rate is likely lower than this.

A brief description of repeaters

I begin by detailing the racial and political profile of likely repeaters compared to respondents who likely did not take the 2018 survey. In terms of partisanship, repeaters are more likely to be strong partisans and less likely to say that they are unsure of their party identification. This pattern is logical since people who are more interested in politics are more likely to agree to take multiple political surveys. Nevertheless, the differences are not particularly large – in the order of 1 to 3 percentage points.

```
                    New Repeater
Strong Democrat   0.260    0.271
Weak Democrat     0.113    0.102
Lean Democrat     0.112    0.099
Independent       0.147    0.142
Lean Republican   0.083    0.090
Weak Republican   0.080    0.086
Strong Republican 0.158    0.193
Not sure/Other    0.047    0.017
```

Repeaters are also less racially diverse than new respondents. For example, 77% of repeat respondents are white compared to 70.6% of new respondents. This reflects the fact that it is more difficult to recruit and maintain people of color on survey panels.

```
           New Repeater
 Asian    0.032    0.024
 Black    0.121    0.096
 Hispanic 0.091    0.068
 Other    0.050    0.042
 White    0.706    0.770
```

Do repeaters affect inferences

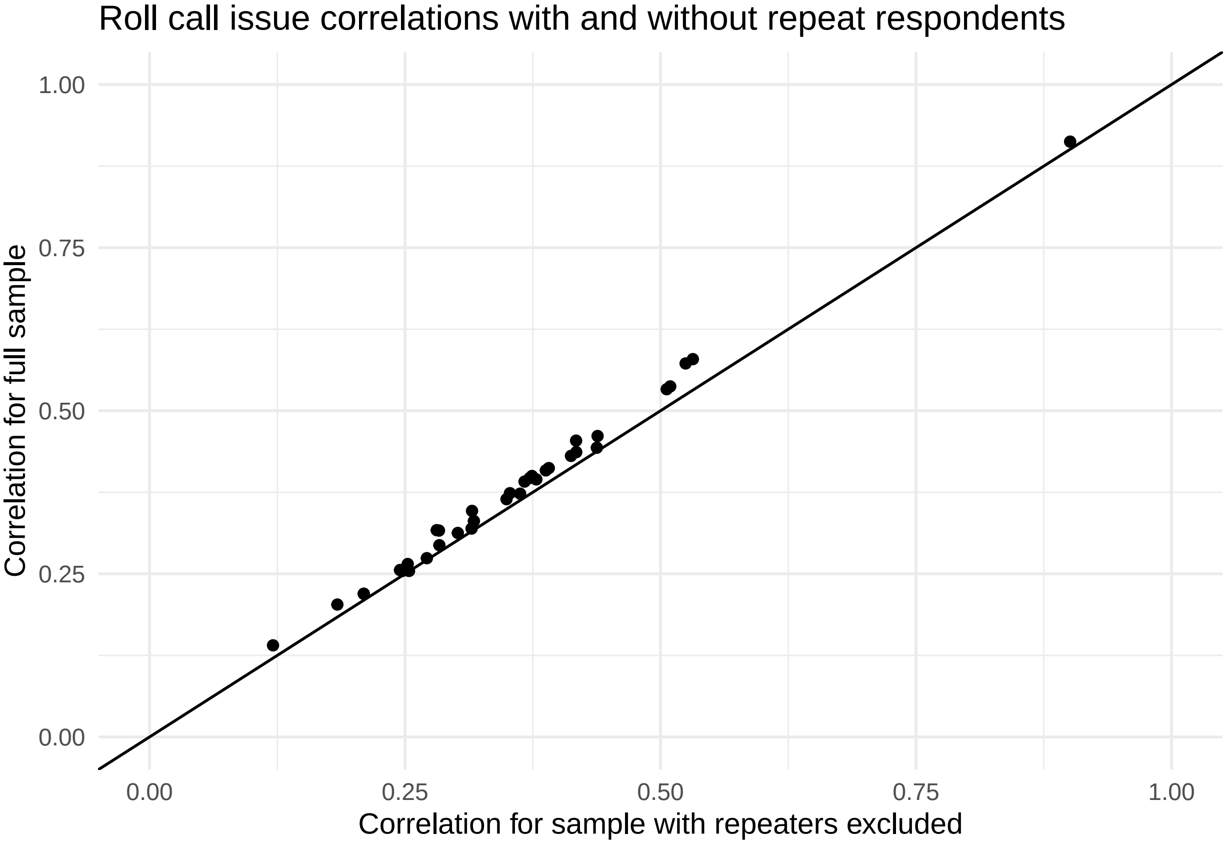
To explore whether the presence of repeat CES takers affects the inferences one is likely to draw from any given cross-sectional dataset, I conduct three analyses: (1) an examination of inter-item correlations on issue questions, (2) a multivariate regression model of vote choice, and (3) an analysis of experimental treatment effects from a conjoint experiment. In each case, I produce one set of estimates using all respondents (new and repeat respondents) in the sample and a second set of estimates after removing those tagged as likely repeaters. Removing likely repeaters has no statistically distinguishable effect on the inferences one would draw from the latter two analyses and only a marginal effect on the first.

Inter-item correlations

One way in which we might expect new and repeat respondents to differ is the extent to which they show cross-item constraint (or structure) to how they answer issue questions. Specifically, taking numerous political surveys may lead respondents to take positions on issue questions that are consistent with how they answer other questions. Constraint is demonstrated by examining how much responses to one question are correlated to responses to other questions. To test this possibility, I compare the correlations in how respondents answered the nine roll call vote items included in the 2020 CES for the full set of respondents and for the sample after dropping likely repeaters.[1] With nine issue questions, there are 36 pairwise correlations to calculate.

The graph that follows is a scatter plot showing each of the 36 issue item correlations calculated and the value that the correlation takes on for all respondents (on the x-axis) and once repeat respondents are dropped (on the y-axis). I have recoded items so that they all point in the same direction, ensuring that all correlations will be positively signed. The 45-degree line is provided as a reference. Correlations falling above this line are stronger in the sample when likely repeaters are included than they are when they are dropped.

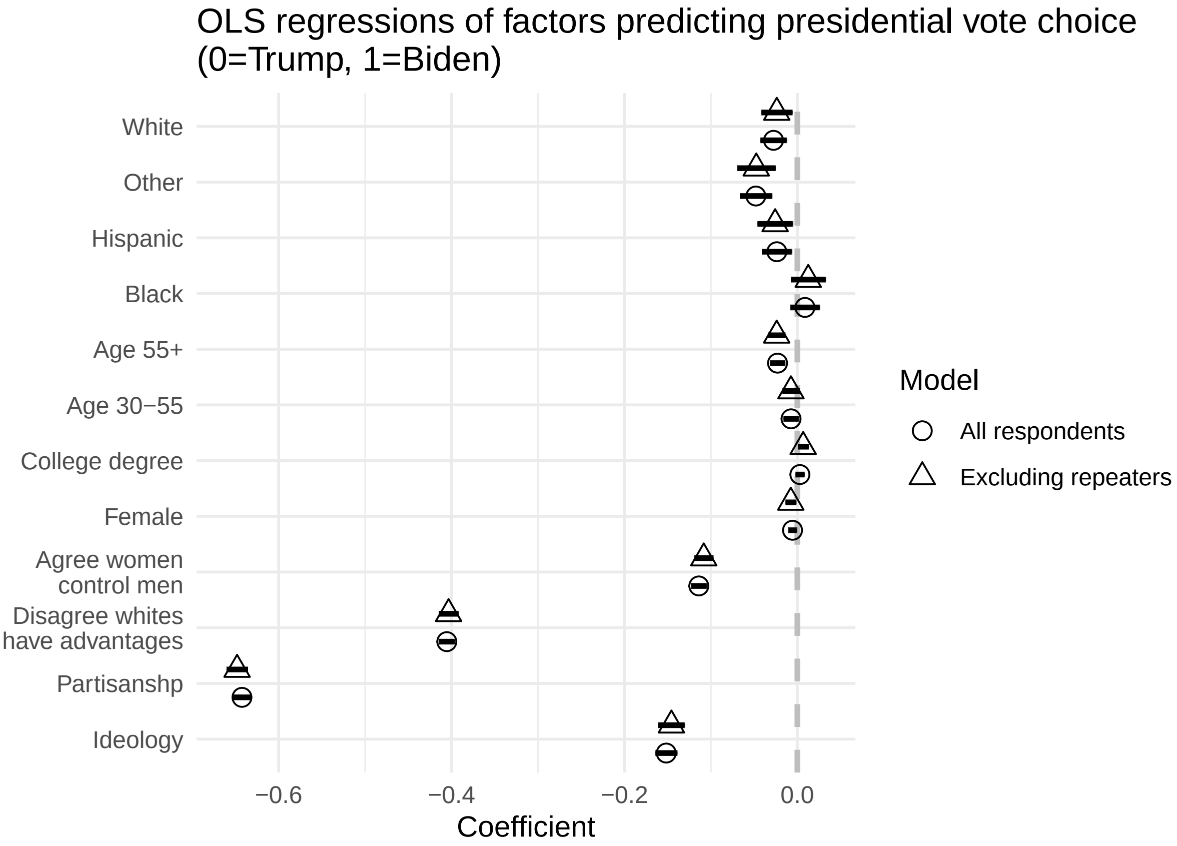
[1] The seven items included questions about (1) amending federal laws to prohibit discrimination on the basis of gender identity and sexual orientation, (2) raising the minimum wage to $15 per hour, (3) confirming Brett Kavanaugh to the Supreme Court, (4) requiring equal pay for men and women, (5) DACA, (6) impeaching Trump for abuse of power, (7) impeaching Trump for obstruction of Congress, (8) the CARES Act, and (9) the HEROES Act.



The plot shows that including repeat respondents in the sample does appear to produce more issue constraint than if such respondents are excluded. Each of the 36 correlation coefficients is larger when likely repeaters are included, though the magnitude of this difference ranges from nearly zero to 0.047. On average, the inter-item correlations for the full sample are .019 larger than when likely repeaters are excluded. Thus, repeat respondents do appear to produce more issue constraint among the full sample, though the effect of their inclusion is modest.

Vote choice model

The figure below plots OLS coefficients from a model predicting whether a 2020 CES respondent said that they voted for Joe Biden (coded 1) or Donald Trump (coded 0) in the 2020 presidential election. The variables were selected according to models of recent election voting behavior (e.g. Schaffner, Nteta, and MacWilliams 2017) and include demographics, partisanship, ideology, and items measuring racial attitudes and sexism. The graph plots the coefficient for each item and the horizontal lines represent 95% confidence intervals. Note that the model that includes the likely repeat respondents produces estimates that are nearly identical to the one using only respondents who are flagged as likely to not be repeaters.



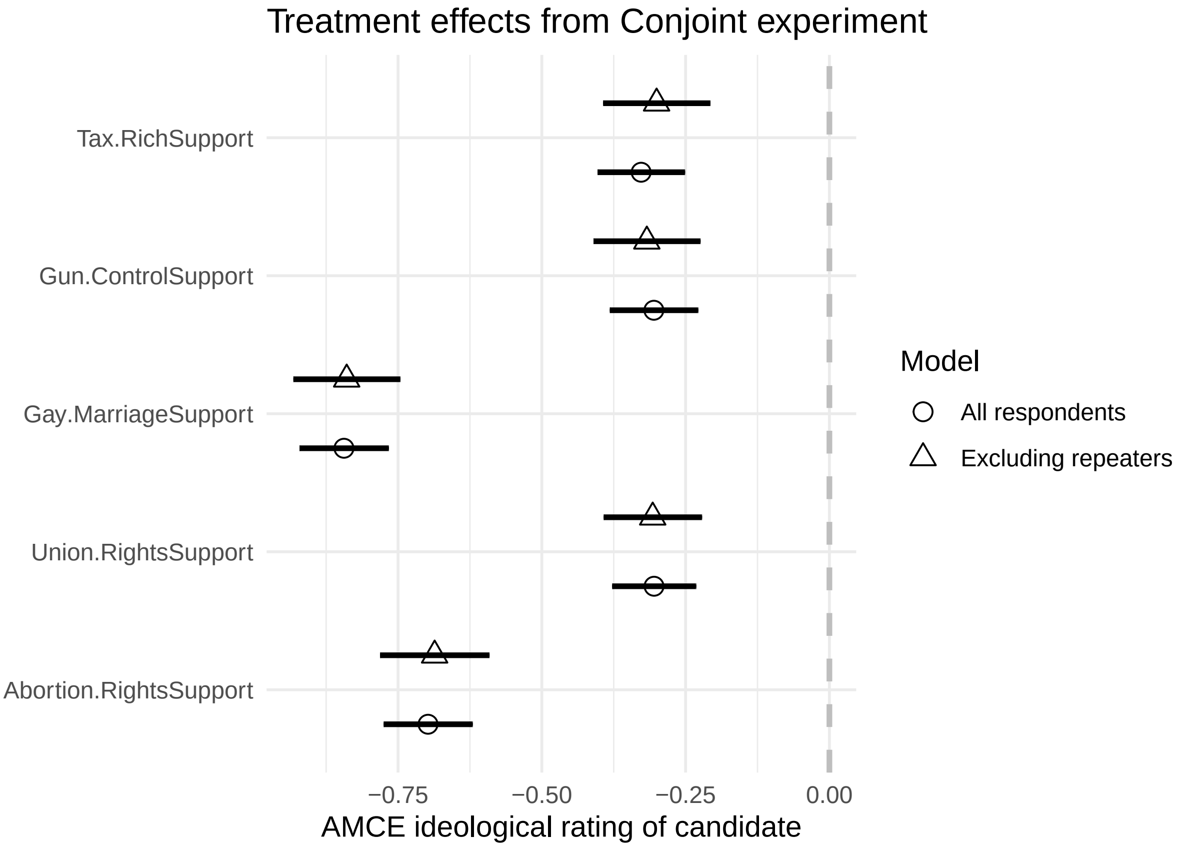
Conjoint experiment

The next analysis relies on a simple conjoint experiment fielded on the Tufts University module of the 2020 CES. In this experiment, respondents were shown a hypothetical candidate's randomly assigned positions on the following issues:

- [Support/oppose] Access to abortions
- [Support/oppose] Protections for workers' right to organize unions
- [Support/oppose] Gay marriage
- [Support/oppose] Gun control
- [Support/oppose] Higher taxes on richest 5% of households

They were then asked to describe this candidate's political views on a five-point ideological scale ranging from Very liberal (1) to Very conservative (5).

The graph that follows plots the AMCEs (treatment effects) for a candidate being assigned a support (versus oppose) position on each issue. Negative treatment effects indicate that candidates who supported a policy were rated as more liberal than those who opposed it. Similarly to the previous analysis, the AMCEs are nearly identical when the analysis is conducted with the entire pool of respondents compared to when repeat respondents are removed.



Conclusion

Because online surveys rely on panelists that take surveys over many years for any particular firm, it is natural that some individuals in a given sample will have taken a previous version of the CES. So far, however, we see little evidence that this has a significant impact on estimates researchers might produce from the data. Repeat respondents are only marginally distinct from "new" respondents when it comes to partisanship. Repeat respondents do demonstrate somewhat higher correlations across issue items, but their inclusion only increases cross-item correlations by a small amount (.019 on average). Inferences drawn from an observational vote choice model and from an experimental task were nearly identical regardless of whether repeat respondents were included or not. Of course, the tests included here are of limited scope, and future reports will test other potential ways that repeat respondents might influence CES surveys. But so far, the results suggest little reason for concern about respondents who take multiple CES surveys.