What Can Be Learned From 500,000 Online Survey Responses About Party Identification?

Alexander Coppock and Donald P. Green*

Columbia University

January 21, 2016

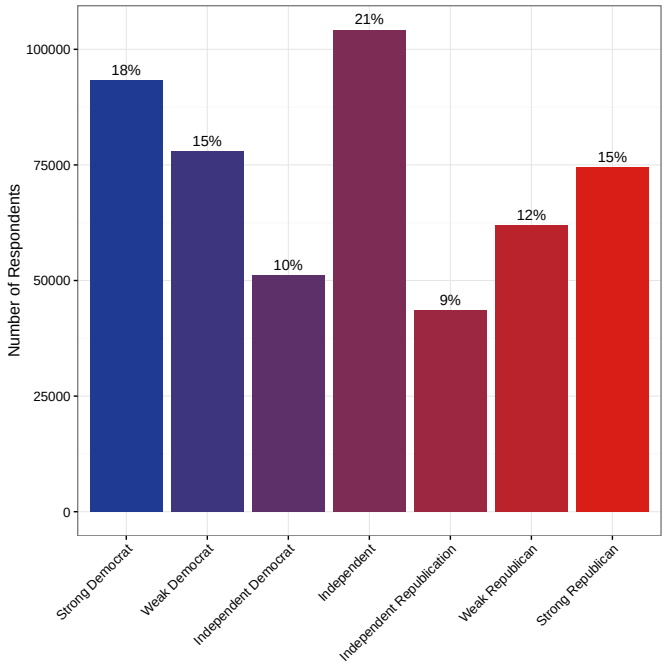
In mid-December 2015, the market research technology firm Lucid added the American National Elections Survey (ANES) party identification question to its battery of standard demographics collected from all survey takers on its marketplace for online survey responses, the Fulcrum Exchange. By December 31st, 2015, Lucid had collected 506,642 individual responses, making this the one of the largest (if not the largest) set of responses to the ANES party ID question ever assembled.

In this document, we assess the utility of these responses for political research. The chief hurdle is that these 500,000 responses are not necessarily representative of any population in particular, because they were not sampled at random from a well-defined population. Undoubtedly, the people who responded to this survey differ from the national population at large. This sample in this study is better educated, has higher incomes, is younger, and is whiter and more female than the national average. A natural question to ask is, are the Democrats and Republicans in this sample like Democrats and Republicans generally, or are they a hopelessly skewed subset? Although Lucid often employs demographic quotas to better approximate specific populations of interest, the data collected for this study did not use this feature. Furthermore, we set out to evaluate the quality of the raw data, so we do not reweight the responses or employ post-stratification to account for the possible differences, measured or unmeasured, between the Lucid sample and the population at large.

The ANES question reads: "Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?" Those who answer with Republican or Democrat are further asked "Would you call yourself a strong [Republican/Democrat] or a not very strong [Republican/Democrat]?" Those who originally describe themselves as an Independent (or other) are asked: "Do you think of yourself as closer to the Republican or Democratic party?" The responses were mapped into a single variable that can take values from 1 to 7, with higher values meaning more Republican. The raw counts of respondents in each of the seven categories are presented in Figure 1.

*Alexander Coppock is a Doctoral Candidate in Political Science at Columbia University. Donald P. Green is Professor of Political Science at Columbia University. We have no financial connection to Lucid and received no compensation for this report.

Figure 1: Distribution of Party ID in Lucid Sample



We first compared our data to existing measures of national partisanship. The NYT/CBS poll has been asking a (non-branching) version of this question for many years. Typically, that organization reports the proportions of Democrats, Republicans, Independents and "Don't Knows." In order to make the datasets comparable, we lumped the Independents and Don't Knows in the NYT/CBS polls together and in the Lucid sample, we considered all respondents who did not pick Democrat or Republican to be Independents.

Figure 2 plots the NYT/CBS time series of partisanship since 2010. Since then, fewer respondents self-identify as being members of either major party. The Lucid estimate is plotted as a single point at the end of 2015. Compared with the CBS survey, the Lucid survey produces a larger proportion of respondents who describe themselves as Democrats or Republicans and a smaller proportion who describe themselves as independents or "other." This figure shows that overall, the Lucid sample is slightly more partisan than the national population. We note, however, that the overstatement of partisanship does not appear to favor one party over the other.

Figure 2: NYT/CBS Partisanship Time Series and One Lucid Partisanship Measure

Next, we compare the Lucid data to national election returns. We aggregated the partisanship responses to both the congressional district level and the county level. Our main summary variable is the proportion of Democrats out of all partisans within a district or a county. We then compared that summary with the share of the two-party vote obtained by the Democrats. For the district-level analysis, this is the share of the two-party vote obtained by the Democratic candidate for Congress in 2012.[1] For the county-level analysis, we used the share of the two-party vote obtained by President Obama.

Figures 3 and 4 present our main results. In both graphs, the Lucid estimate of the Democratic share of partisans (Ds / (Ds + Rs)) is presented on the horizontal axis, with the Democratic electoral advantage (Votes for Ds/(Votes for Ds + Votes for Rs)) plotted on the vertical axis. The points are sized in proportion to the square root of the number of people who responded within each geographical unit. The points are also colored so that more Democratic places are bluer.

A number of features of these plots stand out. First, there is a strong association between partisanship as measured by Lucid and the recorded two-party vote share. The correlation is 0.86 at the district level and 0.51 at the county level. We reiterate that no statistical adjustments have been made – these are the raw proportions. Presumably one could do better by reweighting the data to match district or county demographics, or equivalently, by post-stratification. Second, we have good reason to think that these correlations are attenuated by measurement error: the fewer observations we have in some geographic districts, the more noise is associated with the estimate. This can be seen clearly on the graphs: points that are larger (i.e., they represent the answers of more respondents) are closer to the 45-degree line. Third, the county-level graph confirms some known political features of the US: smaller counties are both more numerous and more Republican, whereas larger counties are fewer in number and more Democratic.

[1] We excluded districts in which the election was not contested by one of the major parties.

Figure 3: 2012 Congressional Elections:
Two-Party Vote Vs. District Partisanship

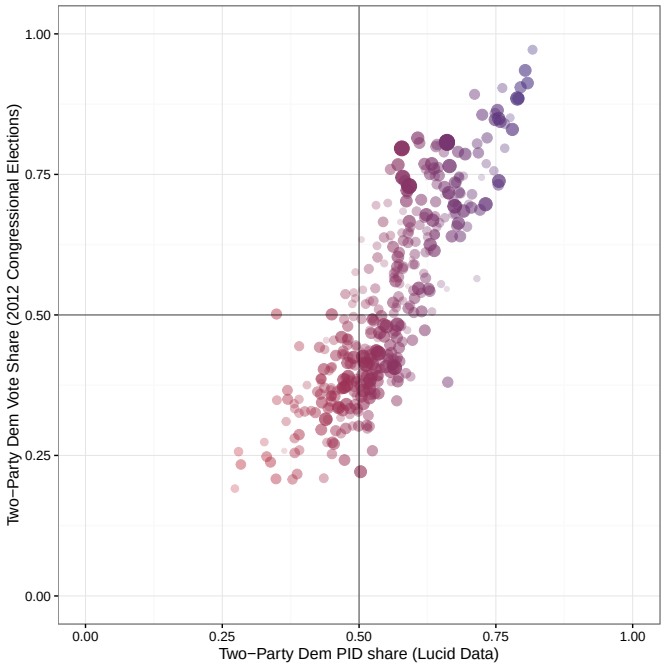


Figure 4: 2012 Presidential Elections:
Two-Party Vote Vs. County Partisanship

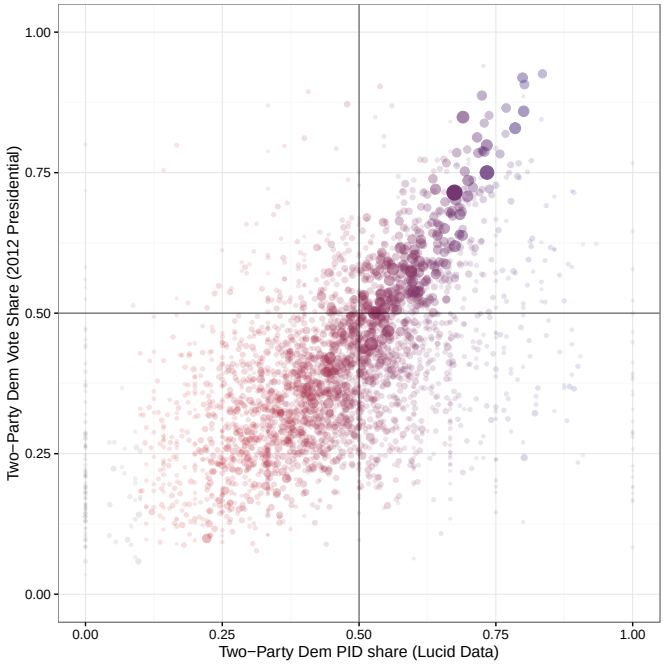


Table 1 shows how the correlation of estimated Democratic partisan advantage with the Democratic electoral advantage increases with sample size. For both the district- and county-level comparisons, we computed the raw correlations for successive tranches of the dataset. Using the only the first 100,000 responses, the correlations are 0.82 and 0.38, at the district and county levels, respectively. These correlations become stronger as we add more data: when we include the full dataset, the correlations grow to 0.86 and 0.51. Presumably, if more data were acquired from those geographic units for which we have relatively small numbers of respondents, then the marginal impact of increased sample size on the correlation with electoral advantage would be even stronger.

Table 1: Correlation of Estimated Partisan Advantage with Electoral Advantage as a Function of Sample Size

	District Level Correlation	County Level Correlation
First 100k	0.8157	0.3802
First 200k	0.8399	0.4720
First 300k	0.8468	0.4902
First 400k	0.8518	0.4888
All 500k	0.8553	0.5068

We conclude from this exercise that the samples traded on the Fulcrum Exchange are useful for political research. First, convenience samples with known party ID can be tapped for survey experiments. Second, the samples can be used for non-experimental work such as election polling. While there of course remain non-statistical sources of uncertainty (e.g., threats of bias due to selection), this exercise has given us confidence that the Lucid sample is on par with traditional samples, that, despite possibly being initially drawn at random from some population, face biases due to non-response. Third, we note that the predictive accuracy of the partisanship variable increases with the number of respondents, suggesting that Lucid will do better and better over time and that, at any given point in time, could increase accuracy through more intensive local recruitment.