Party Identification Weighting - Experiments To Improve Survey Quality

Olena Kaminska and Christopher Barnes
University of Connecticut and Center for Survey Research and Analysis, USA

Key words: weighting, party identification, pre-election polling methodology

1. Introduction

Weighting the pre-election polls by PIDentification (PID) has become one of the most discussed and controversial issues in survey research field. While most studies on the topic examine the stability of party ID and on this basis attempt to prove or disprove the validity of weighting by PID, it may be reasonable to weight by PID even if the distribution of PID is not completely stable. For example, a portion of change in PID may be an indicator of the change in response willingness rather than actual change in PID. If PID is correlated with response willingness, weighting can improve the response rate error.

The purpose of this paper is to present three experiments, which evaluate if weighting by PID can improve the quality of pre-election survey. The experiments were designed to answer four different questions: first, if weighting by PID shifts the vote choice results closer to election outcome, second, what design for PID weights is best, third, how weighting by PID works within the likely voter model and within all population data set and, fourth, if weighting by PID improves the quality of pre-election survey overall, not only vote choice variables. The study is based on 2000 pre-election polls sponsored by the following organizations: ABC News/Washington Post, Gallup/CNN/USA Today, NBC/WSJ, PSRA/Pew and Newsweek.

2. Weighting by PID

2.1 Discussion

As noted, there has been extensive debate regarding PID weighting. Opponents of PID weighting have cited three main reasons against weighting (Blumenthal, 2004): (1) the question about PID is among the last ones in the survey, (2) weighting by PID means using four-year-old exit poll results and (3) PID is not stable.

In a typical pre-election poll the party identification question is usually asked in the demographic section at the end of the questionnaire. Because questions on political preference and issues precede the PID question, they might influence the party identification response. This problem may be exacerbated if different questions or different question order are used in the exit poll (or other data source for weighting).

Using four-year-old exit poll results is slightly different from using four-year-old Census results for demographic variables, given that PID is not a factual measure, but is an attitudinal measure. In contrast to other variables such as gender, age, race, education or income, usually used for weighting, party identification might change as a result of political or social events. Therefore using four-year-old PID distribution might not improve the accuracy of pre-election polls if the PID outcome is different in following elections.

A critical question is: are the shifts in PID that occur in pre-election polls a correct measure of the shifts in population or are they due to survey error? This question is usually answered by argument about stability of party identification. The main reasoning of researchers who decide to weight by PID is that PID is stable over years (Bowers, 2004). Opponents of weighting by PID argue that while party affiliation changes slowly over long-term periods, it can have daily or weekly shifts in pre-election period (Traugott, 2001; NAES press release 2004).

While discussing the stability of PID is important, it is not the only possible explanation of PID shifts in polls. Before making conclusions researchers should account for two other significant issues, which can cause shifts: response rate and likely voter models.

The issue with response-rate is that PID can be related to response willingness. While the shift in PID occurs in polls, it may be due to the shift in willingness of different parties' supporters to participate in surveys. Republicans or Democrats might be more willing to answer pre-election poll questions if their candidate wins the debate as they may be more enthusiastic and have a stronger desire to express that opinion.

The problem with likely voter models is that the model selects "interested" respondents. As Rich Morin notes: "… more members of one party may become interested in the presidential race at any given moment (and therefore qualify for the likely voter pool) while members of other party become momentary bored, distracted or annoyed (and thus are judged less likely to vote)" (Morin, 2000).

Chris Bowers (2004) has compared 12 polls, 6 of which are weighted by PID and 6 others are not. The polls took place during approximately the same time periods and were also compared to next in time polls within each company. Three major conclusions were derived from the experiment:

- Polls weighted by PID show lower variance in vote choice over time than those not weighted by PID.

- The shift over time within each company was higher for those companies that do not weight by PID.
- If the not weighted by PID data sets were weighted by PID, they would look almost exactly like the six weighted polls (the vote choice results would have smaller variance).

This paper presents further research, which evaluate, if weighting by PID improves survey quality.

2.2 Experiment 1

The main purpose of the first experiment is to compare PID weighted vote choice results to those, weighted by original weight. This experiment is conducted using likely voter model. It also addresses the question about the best PID design, comparing two-level (Democrat vs. Republican) to three-level (Democrat vs. Republican vs. Independent) PID weights.

2.2.1 Data Sets

All three experiments were conducted, using next five pre-election surveys: ABC News/Washington Post Poll # 15762: Post Presidential Debate Poll [October 12-15, 2000]; Gallup/CNN/USA Today Poll # 38: Final election Tracking Poll [November 5-6, 2000]; NBC/WSJ Poll # 6010 [November 3-5, 2000]; PSRA/Pew # 10MID: Mid-October 2000 Political Survey [October 18-22, 2000] and PSRA/Newsweek Poll # 2000-NW33: The Homestretch [October 31-November 2, 2000]. The data sets were selected on the basis of two main criteria: the availability of data in the Roper Center for Public Opinion Research archive and the dates the poll was conducted (the closest available to presidential election in 2000, which was held on the 7th of November 2000).

2.2.2 Weighting

Each data set, used in this experiment, has demographic-based weights and likely voter weights. The purpose of the first weight is to adjust the results according to US Census demographics. The likely voter weight is used to exclude respondents who are not likely to vote. The combined weight will be called original weight in this paper.

For each dataset two different weights were created: two-level PID weight and three-level PID weight. They were calculated independently for each poll. PID distribution was taken from 1996 exit poll, according to which 39% of voters identified themselves as Democrats, 35% - as Republicans and 26% as Independent. For the two-level weight only Democrats vs. Republicans distribution was accounted for. The distribution from polls was compared to the exit poll outcome with 52.7% Democrats and 47.3% Republicans. Independent and others received weight equal to one. For the three-level weight Democrats, Republican and Independent distribution was used. "Other party" category and responses, which were coded as "Don't know" or "Refused" received weight equal to 1.

For two-level PID weight the highest value was 1.03 and the lowest - 0.96, which means that Democrat-Republican distribution in these polls were very close to the distribution in exit poll and could not change the results much. The highest weight for three-level PID (step 5) was 1.19 and the lowest – 0.70, which is higher than for two-level weight, but still is considerably low in comparison to typical demographic weights.

Party ID based weight was multiplied by original weight to get PID final weight. Thus PID final weight selects from the data set only likely voters, improves the distribution of demographic variables, bringing them closer to Census data, and improves the distribution of PID according to 1996 exit poll distribution.

2.2.3 Error Calculation

There are different ways to evaluate the accuracy of pre-election polls. The detailed information about 8 methods of evaluation, developed by SSRC in 1949, their advantages and disadvantages can be found in "Review: Was 1996 a Worse Year for Polls than 1948?" written by Mitofsky (1998; 1999). Three of these will be used here: method 3, method 5 and method 7 (it has become traditional to name the methods by numbers in the SSRC list).

The methods used in this analysis:

1. Method 3 is the average error, calculated as the average (without regard to sign) of the percentage point deviation for each candidate between his/her estimate and the actual vote
2. Method 5 is a margin error, which is calculated by subtraction the margin between the top two candidates in a poll from the margin between the same candidates in the election.
3. Method 7 is the chi-square, which tests if the difference between poll and election results is statistically significant.

There is a conceptual difference between method 3 and method 5. Method 5 evaluates the margin between two leading candidates. In other words it compares if the difference between two leading candidates was predicted correct. Method 3 evaluates how close, on average, the percentage for each candidate in poll was to the percentage in elections. There is discussion about how many candidates should be included in method three. One line of the reasoning, which is used here, is to include all the candidates reported by the pollster. Thus method 3 will take into account four candidates from 2000 presidential elections: Al Gore, George W. Bush, Pat Buchanan and Ralph Nader, while method 5 will take into account the Al Gore - George W. Bush margin.

In the actual presidential election in 2000 48.38% voted for Al Gore, 47.87% for George W. Bush, 0.42%

for Pat Buchanan and 2.74% for Ralph Nader. The poll results were compared to these percentages.

2.2.4 Results

As was mentioned earlier, party distribution after weighting by original weight is close to 1996 exit poll for all polls in the experiment. While this is true, Democrats were slightly underrepresented in all surveys. This resulted in higher weights for Democrats as compared to the weights for Republicans. Thus weighting by two-level PID has increased the percentage of Al Gore supporters and decreased the percentage of George W. Bush supporters for each poll (table 1). In all the cases the change was not higher than 1.05%. The outcome for three-level weight is very similar. According to method 7 there is no statistically significant difference between vote choice results weighted by original weight and those weighted by two-level or three-level final PID weight for any poll.

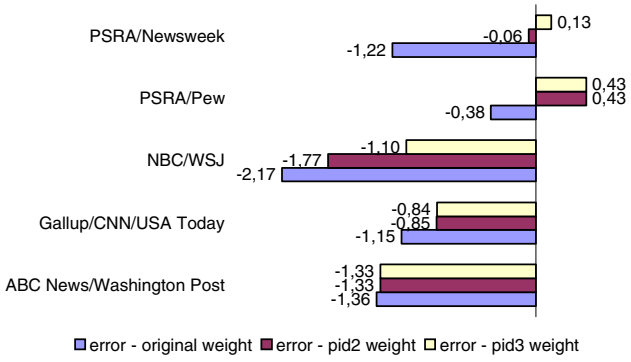
Table1: Vote choice outcome after weighting by two-level PID final weight

	ABC News/Wash Post		Gallup/CNN/USA Today		NBC/WSJ		PSRA/Pew		PSRA/Newsweek	
	original weight	PID final weight	original weight	PID final weight	original weight	PID final weight	original weight	PID final weight	original weight	PID final weight
Al Gore	44.91	44.94	45.07	45.34	44	44.36	44.84	45.55	43.28	44.33
George W. Bush	46.92	46.89	46.69	46.41	47.48	47.12	45.05	44.3	44.97	43.97
Pat Buchanan	0.57	0.57	0.65	0.66	1.56	1.55	0.62	0.62	0.12	0.12
Ralph Nader	3.6	3.6	3.74	3.74	3.54	3.55	4.09	4.1	4.94	4.96
LV counts	1472		2350		1026		663		808	

According to election outcome Al Gore won the popular vote. As can be referred from the table 1, none of the polls predicted the right winner before weighting and two of them, PSRA/Pew and PSRA/ Newsweek polls, could predict the right winner after weighting by PID. Given this, although weighting by PID has increased the margin error for PSRA/Pew poll (regardless of the sign), in the same time it improved the results by predicting the right winner.

One of the main issues of interest is directional: does weighting by PID shift results closer to election outcome? Margin errors (method 5) decreased for four out of five polls after weighting by PID. This means that all the polls, except for PSRA/Pew poll, predict the difference between two leading candidates better after weighting by PID. Figure 1 presents the results for margin errors before and after weighting by PID within each company.

Figure 1: Comparison of margin errors for weighted by original weight, by two-level PID (pid2) weight and by three-level PID (pid3) weight poll results.

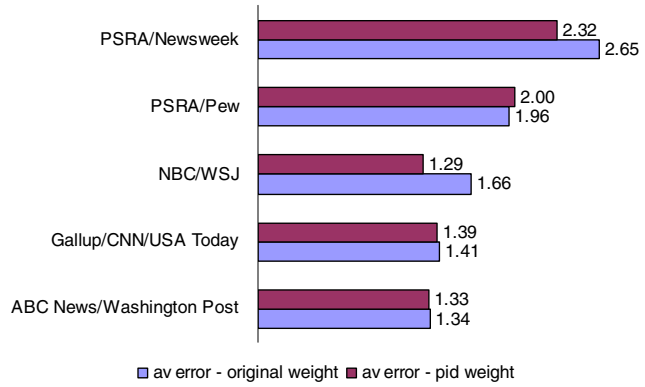


According to binomial probability theory, there is 19% of chance that four or more out of five data sets can improve the results after weighting. This is too low for rejecting the hypothesis that improving does not occur. Considering that only five data sets were used this result suggests that further experiments should be conducted.

Both two-level and three-level PID weights show very similar results and there is no enough evidence to claim that either is better.

There was no or almost no change in the average error (method 3) for all five polls after weighting by PID. The error for four polls decreased after weighting by three-level PID and for PSRA/Pew poll - slightly increased (figure 2). The difference between errors for most companies is too small to make strong conclusions about the one-directional shift.

Figure 2: Comparison of average errors for weighted by original weight and by three-level PID weight poll results



2.3 Experiment 2

The main idea of experiment 2 is to examine if PID weight improves the outcome for all the population, including unregistered and "unlikely" to vote respondents. Thus likely voter model was not used for this experiment. Here the vote choice outcome was calculated for unweighted, weighted only by PID and weighted by demographic weight data sets. Poll outcomes were compared to NES 2000 vote choice distribution.

For the second and third experiments same data sets were used, except for NBC/WSJ Poll. The reason for this was that NBC/WSJ poll did not include non registered respondents. Thus comparing the results for this poll to NES would not be theoretically correct.

2.3.1 Weighting

The vote choice frequencies were run for three following weighting designs:

Unweighted – the data set was simply unweighted and all the cases were included.

Weighted by demographic weight – each data set has a developed by the company demographic weight. Different companies might use different demographic factors for developing weights, which depended only on companies' choice.

Weighted by "PID only weight". This weight calculation was different from the one for experiment 1. The poll PID distribution was derived from unweighted data set. This distribution was simply compared to weighted NES PID distribution. Weighting in NES does not have PID factor but accounts for demographics and selection probability (Burns, Kinder, Rosenstone, Sapiro, 2002). Only the first question about PID affiliation without following questions about leaning was used. Only Democrat-Republican distribution was accounted, excluding Independent. To calculate the only PID weight, NES outcome was divided into the outcome from the poll, both added to 100% beforehand.

2.3.2 Results

According to chi-square test, there was significant shift in vote choice results after weighting by PID in comparison to unweighted results for ABC News/Washington Post, PSRA/Pew and Gallup/CNN/USA polls. The shift after weighting by PID was also significant if compared to weighted by demographic weight results for PSRA/Pew and Gallup/CNN/USA polls. As demonstrated in figure 3, the margin error has decreased for all four data sets. According to binominal probability theory, there is only 6.25% chance that weighting by PID will improve the outcome for four out of four polls.

Figure 3: Comparison of margin error for unweighted, weighted only by PID and weighted by demographic weight polls results

Error, calculated by method 3, did not show one-directional results. For ABC News/Washington Post and NBC/WSJ polls the error was smaller for weighted only by PID weight as compared to unweighted and weighted by demographic weight outcome.

For the Gallup/CNN/USA Today poll, the error after PID weight was in between other two, and for the PSRA/Pew poll it was the highest. While the error differences within each poll were small and were not one-directional, no strong conclusions can be made about the influence of weighting by PID on average error.

2.4 Experiment 3

The main idea of experiment 3 is to examine if weighting by PID improves the outcome for other variables, not related to vote choice. For this purpose demographic variables, such as gender, age and race, were chosen, as their distribution in the population is known from the Census.

For this experiment likely voter model was not used and all respondents were included. The comparison was done between the outcomes for unweighted and for weighted only by PID data. The error was calculated as a difference between poll outcome (unweighted and weighted by PID) and Census.

Each variable was divided into two groups:

- Gender: female and male
- Age: "18 to 44 years old" and "45 years old and above".
- Race: "White" and "Black / African-American"

2.4.1 Results

There was no one-directional change in error for gender and age after weighting by PID. For gender error decreased in two polls, increased for two others and did not change in one. For age the errors for unweighted and weighted by PID data were very close to each other. For race in all 4 polls the error decreased after weighting by

PID. The change in all cases was too small to legitimize any strong conclusion.

The chi-square test showed no significant difference between unweighted and weighted by PID outcome. Thus weighting by PID does not seem to change the results for demographic variables. It does not seem to improve constantly the demographic distribution, but importantly it does not seem to violate it also.

2.5 Conclusion

Weighting by PIDentification is one of the most controversial topics in survey methodology. In this paper we would like to shift the debate from the stability of PID, as reasoning for weighting (or not weighting) by PID, to experimental research designed to measure the performance of PID weighting. The results of these initial experiments indicate that there is potential for PID weighting to be an addition to likely voter models as aids to election predictions. The results are not definitive by any means, rather suggestive of the need for further exploration.

Four main conclusions can be made from our three experiments:

- Weighting by PID seems to improve prediction for the margin between two leading candidates. This is true for likely voter data (both two-level and three-level PID weights), the same as for the whole population data sets.
- Weighting by PID does not seem to have one-directional influence on average error for four presidential candidates. It does not seem to improve, but it also does not violate, the overall vote choice results.
- There is no substantial difference between two-level PID and three-level PID weight designs.
- PID weight does not seem to change the demographic variable distribution.

The attention should be paid to the fact that this paper experiments are based on the polls from one year and one presidential election – November 2000. The exit poll PID distribution in 1996 and 2000 years were very close: 39% Democrats, 35% Republican and 26% in 1996, 27% in 2000 Independent. Careful consideration and other research should be conducted before the results can be extrapolated on other time points.

References

Annenberg Public Policy Center of the University of Pennsylvania (2004). Party Identification Shifts toward Republicans, but Democrats Still Lead. Annenberg Data Show: Press Release November 19, 2004. The Annenberg Public Policy Center of the University of Pennsylvania. 19 November 2004. <http://www.annenbergpublicpolicycenter.org/naes/>.

Blumenthal, Mark "Weighting by Party". September 23, 2004.<http://mysterypollster.typepad.com/main/2004/09/weighting_by_pa.html>

Bowers, Chris "Rapid Poll Movement is a General Election Myth". 18 September 2004. <http://www.mydd.com/story/2004/9/18/454/00913>.

Burns, Nancy, Donald R. Kinder, Steven J. Rosenstone, Virginia Sapiro, and the National Election Studies. American National Election Study, 2000: pre- and post-election survey [Computer file]. 2nd ICPSR version. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer], 2001. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002.

Mitofsky, Warren J. Reply to Panagakis the Public Opinion Quarterly, Vol. 63, No. 2. (Summer, 1999), pp. 282-284.

Mitofsky, Warren J. Review: Was 1996 a Worse Year for Polls Than 1948? The Public Opinion Quarterly, Vol. 62, No. 2. (Summer, 1998), pp. 230-249.

Morin, Richard. "Unconventional Wisdom; New Facts and Hot Stats from the Social Sciences;" The Washington Post. Washington D.C.: Oct 15, 2000. Pg B.05

Traugott, Michael W, Assessing Poll Performance in the 2000 Campaign, The Public Opinion Quarterly, 2001, 65, 3, fall, 389-419