# A Method for Weighting Survey Samples of Low-Incidence Voters

In this paper we describe a method for weighting surveys of a sub-sample of voters. We focus on the case of Latino voters. And we analyze data for three surveys: two opinion polls leading up to the 2004 presidential election, and the national exit poll from the 2004 election. We take advantage of much data when it is available, the large amount of data describing the demographics of Hispanic *citizens*. And we combine this with a model of turnout of those citizens to improve our estimate of the demographics characteristics of Hispanic *voters*. We show that alternate weighting schemes can substantively alter inferences about population parameters.

[This is an incomplete version of the paper, it omits calculations of uncertainty which are some of the fundamental quantities of interest of the paper.]

July 19, 2005

**R. Michael Alvarez**
**California Institute of Technology**

**Jonathan Nagler**
**New York University**

Paper prepared for the 2005 Political Methodology Summer Meeting, Tallahassee, Florida, July, 2005. The authors can be reached at rma@hss.caltech.edu, and jonathan.nagler@nyu.edu. We thank Stan Greenberg, Mark Mellman Jim Gerstein, and Matt Hogan for stimulating discussions about this topic. Last, we thank our colleagues Marisa Abrajano and Lisa Garcia Bedolla for their work with us on related projects. Error and omissions are our responsibility.

# 1 Introduction: Sampling and Weighting Matter

In recent elections, studying Hispanic political behavior has become a subject of intense activity, both among academics and practitioners.[1] Partly this interest is due to the growth in the size of the Hispanic population, which the U.S. Census Bureau estimated to be 37.4 million in 2002, about 13.3 percent of the nation's population (U.S. Census Bureau 2003). While the Census Bureau has estimated that perhaps as many as four of ten Hispanics are not citizens in the voting-aged population (and hence are not eligible to vote), 45% of the Hispanic citizen voting-aged population is estimated to have participated in the 2000 election, and almost 79% of the registered citizen voting-aged population voted in 2000.[2] This interest in Hispanic voting behavior has also arisen because of the concentration of Hispanic voters in certain "battleground" states like Florida, New Mexico, Arizona and Nevada; also, there are some who have argued that Hispanic voters, due to their religious affiliations, cultural concerns and partisanship, might be a critical campaign target for Republican and Democratic strategists (Alvarez and Garcia Bedolla 2003).

But despite the interest in Hispanic political behavior, it is often difficult and expensive to collect extensive surveys to study their political attitudes and behavior. As Hispanic voters constitute a small fraction of the national electorate but are geographically concentrated, a straightforward approach towards generating a sample of Hispanic voters using traditional random digit dialing (RDD) techniques (for example, calling randomly generated telephone numbers, then using question filters to reduce the sample to only those self-identifying as Hispanic and who claim to be registered voters) could be very expensive, requiring a the termination of many interviews during the filtering process.[3]

The cost and difficulty of generating samples of Hispanic voters using traditional RDD methodologies has led to the use of list-assisted telephone interviewing. One approach ("Listed Hispanic Surname Sampling") utilizes households with listed telephone numbers,

but only after extracting from this list only those households with Hispanic surnames, using a database of Hispanic surnames. This approach has two potential coverage errors, one due to the use of listed households (thus not including unlisted households or households who have recently changed their number) and the other due to possible inaccuracies in the use of a surname database, and requires filters to reduce the sample to voters. The other approach ("Voter File Surname Sampling") uses another list, most commonly a voter registration list, and extracts from that list only those with Hispanic surnames. The "Voter File Surname Sampling" approach requires the least amount of filtering, but again relies on the accuracy of the voter registration list itself as well as the Hispanic surname database. But these latter two types of list-assisted telephone interviewing are less costly than the various RDD techniques discussed earlier.[4]

Of course, it is possible for researchers to use combinations of each of these designs, for example, generating two independent samples (one RDD based, one based on surname or registration lists, combining them into a "dual frame" sampling design). Provided that the sample sizes are sufficiently large, this approach could allow for more detailed study of the precise differences of each sampling approach. There is also the possibility that by using a dual frame design, researchers can leverage the strengths and mitigate the weaknesses of each respective sampling approach, though more study of dual frame designs is necessary (see Mitofsky et al. 2005, for a recent example of such a study).

Second, once the sample has been collected, researchers must then develop a weighting algorithm to adjust the sample, as necessary, so that it corresponds to known population parameters (typically age and education), or that it adjusts for design characteristics of the sampling approach. The development of sample weights, though, is typically not given much discussion in the reporting of survey results, even though how weights are constructed and applied may play a important role in the substantive interpretation of survey results. While true in the context of more typical national population surveys, the development and

implementation of weighting algorithms might be particularly important in the context of sampling populations like Hispanic voters.

In the end, accurate inference from a sample can hinge on a well-constructed and accurate weighting methodology. If the known and observable attributes (say age, which we observe in the population and measure in the sample) are correlated with the political behavior we want to estimate (say the Kerry-Bush vote in the 2004 election), then if we have a sample that does not accurately represent the age distribution of the electorate adjustment to the age distribution of our sample can help improve the accuracy of our behavioral estimate. This is where reliable weighting methodologies can help improve the accuracy of our inferences; this is also where poor weighting methodologies can lead our inferences astray.

In this paper we quickly survey the issues associated with surveying low-incidence populations like Hispanic voters. We then develop a weighting procedure which we apply to some samples from the controversial 2004 presidential election. We conclude with a discussion of future work.

## 2 The 2004 Presidential Election: Conflicting Surveys

In the 2004 presidential election, these issues about studying Hispanic political behavior loomed large. Before the election, it was widely known that the Bush reelection effort had targeted an increase of 3 to 5% in their Hispanic vote share as critical for their efforts, in an attempt to raise the Republican share of the presidential vote from 35% (in 2000) to at least 38 to 40%.[5] Democrats, feeling this pressure on what they also knew was a vital component of their efforts to unseat President Bush, unveiled in July 2004 a $1,000,000 Spanish-language advertising campaign, the largest such Democratic presidential effort.

Hispanic voter surveys conducted early in the 2004 presidential election cycle, before the Democratic National Convention in July, showed President Bush as lagging his target of 35% of the Hispanic vote. A poll conduced in February 2004 by Democracy Corps showed Bush with 34% of the Hispanic vote, and another Democracy Corps Hispanic voter poll (this one with a different sampling methodology) showed Bush with only 30% of this national Hispanic vote. Other polls taken throughout the 2004 presidential election cycle consistently showed Bush with roughly a third of the Hispanic vote or less, or as Leal et al. (2005) computed, an average of 32%.

The shock came when the first wave of complete exit poll results came out. One set of exit poll results, those from the Edison/Mitovsky "National Election Pool" (NEP) national exit poll, estimated the Bush Hispanic vote share at 44%. Given that the NEP exit poll numbers were subscribed to by most of the major news organizations, like CNN and the Associated Press, these estimates were widely reported in the days following the presidential election and were commonly accepted by many. Another set of exit poll estimates was released by the Los Angeles Times, which in an independent national exit poll effort estimated the Bush Hispanic vote share at an even more surprising 45%. A last exit poll effort was conducted by the Willie C. Velazquez Research Institute (WCVI), and their estimates served to deepen the controversy, as their Bush Hispanic vote estimate was very much in line with the estimates from all the surveys before the election, giving Bush slightly over 31% of the vote (Leal et al. 2005).

The discrepancy between two major media organization exit polls, WCVI's and the pre-election Hispanic polls has already been discussed elsewhere (Leal et al. 2005), and is not our direct concern in this paper. It is of course possible that the major media exit polls were more accurate than the pre-election telephone surveys, either because the latter used sampling designs that did not draw representative samples from the Hispanic electorate or because their weighting schemes were incorrect. We are somewhat reluctant to accept that

conclusion without much further analysis than has been conducted to date; first, there were at least ten pre-election polls whose Bush Hispanic vote estimates are highly correlated, and these polls used different sampling and weighting schemes; second, because the WCVI's exit poll estimate does vary from the two media exit poll estimates substantially.

Rather, the motivation for our analysis here is to develop and analyze algorithms for weighting Hispanic voter samples. Our hope is to shed light on the issue of appropriate weighting schemes for these types of samples, assuming of course that the sampling approaches used to gather data from Hispanic voters (or other low-incidence populations) are appropriate. Having better sample weighting algorithms should help insure more accurate inferences in the future, which not only will help with post-election interpretations of voting behavior but which should help insure more accurate assessments of the preferences of this vital and growing segment of the American electorate for policymakers.

# 3 Designing Weights - Goal

We want a method of weighting survey samples of low-incidence *respondents* to enable us to draw inferences about the opinions of the population of *voters* within the low-incidence population, which is a subset of the population of citizens who are members of the minority population. We will refer to **voters** and **citizens** to indicate the distinct groups: citizens being the population, voters being the sub-population we wish to target. Standard weighting algorithms attempt to match samples to populations based on observed attributes of a census of the population (Levy and Lemishow 1999). With minority respondents this is made more difficult because the characteristics of the population of voters, as opposed to the population of persons (or citizens in our case), is *not known*. Thus we face a fundamental problem: we want to weight a sample to be representative — on some set of observable characteristics — of a population, but we do not know the characteristics of the population.

Here is the problem stated more precisely. We sample a set of minority voters at time $t$. We observe the distribution of responses $\hat{\Omega}$ to a question. We would like to produce an estimate of what the true value of $\Omega$ is in the sub-population, and provide a standard error about that estimate. Obviously if we were confident that our respondents are drawn at random from the sub-population we are trying to sample, then this is a solved problem. However, for many reasons the sampling process might not be random (Republicans may not like to answer the phone; young people may not be home to answer the phone; etc). Thus after obtaining a sample, we want to produce weights which when multiplied by $\hat{\Omega}$ would give us an unbiased estimate of $\Omega$.

This is not a 'new' problem: survey shops have been computing and providing weights with surveys for a long time, especially for surveys of political behavior. We first describe a simple algorithm for producing weights, then explain our technique for dealing with our problem of unobservable population characteristics.

The first decision one faces in computing weights is what observable factors to weight to? We want to choose observable characteristics that would predict the political variable of interest. So, if we are interested in correctly estimating the proportion of voters who will vote for Kerry over Bush (the problem in our examples above of the 2004 presidential election), we want to weight our sample to match the characteristics of voters along a set of dimensions that would accurately predict voting behavior. A necessary condition for the characteristics to be useful as weights is that they must be known at the population level, and they must be known for the sample respondents. Many surveys weight the respondents to match the age and education profile of the population.[6] If accurately computed such weights would no doubt improve the estimate of the Kerry vote at the population level, respondents' age and education are predictors of the vote. However, obviously there could still be much error remaining. We improve on this somewhat in this paper by weighting to four characteristics: age, education, sex, and marital-status. There is a practical reason we do this: the weights

are intended to correct errors made in sampling. Errors in sampling are not going to be random. We might find that better educated persons are less likely to end up in our sample because they have unlisted phone numbers, and we might find that single people are less likely to end up in our sample because they are not at home as often as married people. Any characteristic that determines success of sampling *and* presidential vote-choice is something that we would want to weight to.

## 3.1 A Simple Discrete Weighting Scheme

For simplicity we describe a weighting scheme based on two characteristics: age and education.[7] We group all respondents into one of five age categories, and one of five education categories. This gives us twenty-five cells that contain all of our respondents. For any respondent, we want to compute a weight corresponding to their cell that is the inverse of the probability that someone in that cell is drawn at random from the population. To compute an population parameter of interest, we would then simply compute the 'weighted-average' of the sample values. Now we define three quantities:

- $\mathbf{SAMP}(i, j)$ = the proportion of the sample that is contained in cell $(i, j)$.
- $\mathbf{PTT}(i, j)$ = the proportion of the population (or, 'proportion of total turnout') that is contained in cell $(i, j)$.
- $\mathbf{W}(i, j)$ = the weight to be applied to each respondent in cell $(i, j)$.

Note that the weight must satisfy:

$$\mathbf{W}(i,j) \ * \ \mathbf{SAMP}(i,j) \ = \ \mathbf{PTT}(i,j) \tag{1}$$

So, the weight will be given by:

$$\mathbf{W}(i,j) \;=\; \mathbf{PTT}(i,j)/\mathbf{SAMP}(i,j) \tag{2}$$

We first note that $\mathbf{SAMP}(i,j)$ is observed, so to compute that is not a problem. However, the quantities $\mathbf{PTT}(i,j)$ are *not* observed. Not only do we not observe the characteristics of Hispanic voters, we do not even necessarily observe the characteristics of the set of eligible Hispanic voters (i.e., Hispanic citizens over the age of 18). So this is the problem we must solve.

# 4 Estimating the Population Characteristics

We are able to observe the demographic characteristics of the population of respondents, but not of *voters* for 2000. This is provided by the decennial census. However, given the rapid change in the Hispanic population, the 2000 population numbers would not be correct for 2004. We can approximate the demographic characteristics of the population of 2004 Hispanic respondents by using the monthly Current Population Survey (CPS). To do this, we sum over the 12 most recent Current Population Surveys leading up to the November 2004 election. The CPS is a monthly survey of approximately 50,000 households conducted by the Census Bureau. It is a high-quality survey with a response rate of over 90%. If there is a gold standard for measuring the demographic characteristics of the U.S. population in the period between the decennial census, the CPS is it.

However, in the CPS we only observe the the characteristics of the sub-population of *voters* every second year via the CPS Voter Supplement in which respondents are asked whether or not they voted, in addition to the standard battery of CPS items. However, we wish to take advantage of the wealth of information about population characteristics available from the other 11 months worth of CPS data. We can do this by producing a

probabilistic model that provides a mapping from citizens to voters. The task at hand is to: 1) produce the best possible model mapping citizens to voters; and 2) compute correct standard errors of the final quantity of interest that take into account both the uncertainty about the model parameters, and the uncertainty produced from the sampling process.

Next, there is a fundamental modeling choice we consider: we can estimate a model that maps all citizens to voters, and assume that minority voters behave according to that model; or we can estimate a model of minority voters. The first model would probably be mis-specified, but we could estimate it very precisely as we have a large sample of citizens. The second model would not necessarily be mis-specified, but we would no doubt estimate it with a large degree of uncertainty as we would be forced to estimate the model on a relatively small sample. In practical terms: the first model would be a fairly standard demographic model of turnout, with a Hispanic dummy variable on the RHS; the second model would be the same model, but without the Hispanic dummy and estimated on just the subsample of Hispanic citizens. In this paper we try to compare those trade-offs empirically, and produce a 'best-practice' method for weighting samples of minority respondents to produce estimates of preferences or behavior of minority voters. We present the results of such a weighting methodology to samples of Hispanic voters, samples drawn using a variety of procedures. We evaluate the effectiveness of our weighting methodology relative to industry practices.

# 5 Estimation

We weight based on four characteristics: marital-status, sex, age, and education. As these characteristics have 2, 2, 5, and 5 possibilities, respectively, we can group our respondents into 100 cells. We are going to use two different methods of estimating weights, and so we describe two parallel procedures below that will yield two estimates of a weight for each cell.

As described above, we determine $\mathbf{PTT}(i, j, k, l)$ by estimating two different models with data from the November, 2000 Current Population Survey. The first model uses all citizens, and estimates the probability of a respondent turning out as a function of age, education, marital-status, gender, and ethnicity (Hispanic or not-Hispanic). We estimate this model via logit, and use the estimated coefficients to assign a probability of voting to a respondent in any of our 100 cells. The second model uses only Hispanic citizens, and estimates the probability of a respondent turnout out as a function of age, education, marital-status,k gender, and ethnicity. We estimate this model via logit, and use the estimated coefficients to assign an alternative probability of voting to a respondent in any of our 100 cells. The results are reported in Table 1. The models perform almost equally well for predicting Hispanic voters: the model based only on Hispanic voters produces two more correct predictions out of 4122 Hispanics than does the model based on all voters.

**[Table 1 Here]**

We can then multiply these estimated probabilities by the proportion of Hispanic citizens in each cell as calculated from the pooled Current Population Surveys, and that gives us the values of $\mathbf{PTT}(i, j, k, l)$, or the proportion of total turnout in each cell. These are the population values that we want to weight to. Weighting to these values is just a question of arithmetic as described above. We can compute the sample values in each cell ($\mathbf{SAMP}(i, j, k, l)$) from each of the surveys we describe below. Then the weights are computed directly from combining $\mathbf{PTT}(i, j, k, l)$ and $\mathbf{SAMP}(i, j, k, l)$.

In Table 2 we give the unweighted and weighted distribution of age and education. Note the major corrections made at the top and bottom of the education distribution to the NEP exit poll.[8] The NEP exit poll education distribution clearly under-represents those at the lowest rungs of the educational attainment scale, but over-represents college graduates and those with post-college degrees.

**[Table 2 Here]**

## 5.1 Hispanic Samples

The February 5-16, 2004 Democracy Corps Hispanic survey was comprised of three samples. The first sample was a national sample of Hispanic likely voters, with 1564 respondents. Second, there was an oversample of 363 non-Cuban Hispanic likely voters from Florida, and a second oversample of 559 Hispanic likely voters from three battleground states with large numbers of Hispanic voters: Nevada, New Mexico and Arizona. The three samples had estimated margins of error or 2.5%, 5.2% and 4.2%, respectively.[9]

In Table 3 we present the unweighted results of this survey for partisanship and Kerry vote, along with the same values weighted by: a) the weight calculated by Democracy Corps; b) the weight calculated using our turnout model based on all respondents; and c) the weight calculated using out turnout model based only on Hispanic voters. We see fairly small differences across any of the weighting schemes.

**[Table 3 Here]**

We next use the July 2004 Democracy Corps Survey of Hispanic voters. The July 14-22, 2004 Democracy Corps Hispanic survey was a national sample of 1,000 Hispanic likely voters from twelve states that together represent over 85% of the Hispanic electorate in the United States: California, Texas, Florida, New York, Arizona, New Mexico, Nevada, Illinois, New Jersey, Ohio, Colorado and Michigan. This sample has an estimated margin of error of 3.1%.[10]

In table 4 we again present the unweighted results of this survey for partisanship and Kerry vote, along with the same values weighted by: a) the weight calculated by Democracy Corps; b) the weight calculated using our turnout model based on all respondents; and c) the weight calculated using out turnout model based only on Hispanic voters. Here we also use weights computed simply by using reported turnout of Hispanics from the November,

2000 CPS. In this case, our weights give a slightly smaller Kerry vote than the Democracy Corps weights.

**[Table 4 Here]**

The 2004 National Election Pool (NEP) national exit poll was comprised of two surveys. One was an exit poll conducted in 250 poll places throughout the nation on election day; 11,719 voters were interviewed as they left the polls on election day. Additionally, the exit poll sample of voters was supplemented by a sample of absentee or early voters in 13 states with high levels of absentee or early voting; 500 (or 2000?) absentee or early voters were interviewed using a pre-election telephone poll.[11]

In Table 5 we present similar results for the 2004 VNS exit polls. We again present the unweighted results of this survey for partisanship and Kerry vote, along with the same values weighted by: a) the weight calculated by VNS; b) the weight calculated using our turnout model based on all respondents; c) the weight calculated using out turnout model based only on Hispanic voters, and d) the weight directly from Hispanic turnout in the November CPS.[12] Here any weighting scheme we employ gives substantially different results than the VNS supplied weights.

**[Table 5 Here]**

# 6 Evaluation and Uncertainty

We have yet to comment on the uncertainty in the weighting scheme above, how to estimate it, how to use it in analyses of survey data, or how to evaluate it. How accurate is our ultimate result: our estimate of the proportion of Latinos who voted for John Kerry? As we pointed out in motivating the problem, the most fundamental uncertainty we face is that we do not know the characteristics of the population we wish to weight to: the set of Latinos

who voted for president in 2004.

The sources of error and/or uncertainty are:

1. The sampling error in determining $\mathbf{CVAP}(i, j)$ for the various cells we use. The total sample we use is approximately 75,000 Latinos, and we are trying to estimate 100 cell values from that.

2. The sampling error from the estimation of our turnout model.

3. Any fundamental uncertainty in our estimates of turnout as we do have a stochastic model. When dealing with a large enough set of voters, this is not a large problem for us. Were we to estimate the parameters of the model very precisely, all we would need is a probabilistic statement of how likely any individual is to vote: there would virtually no uncertainty left over the aggregate results for millions of voters.

4. Error from incorrect model specification. We are not sure we have the correct model of the determinants of voter turnout.

For several of these quantities we can compute the distribution of the error. Obviously the distribution of the sampling error in determining $\mathbf{CVAP}(i, j)$ is known. And we can compute the distribution of the sampling uncertainty about our estimation of $\mathbf{Pr} - \mathbf{Vote}(i, j)$ via now commonly used simulation methods. In order to deal with uncertainty over model specification, we can adopt a model averaging framework and produce a posterior distribution over outcomes (Bartels 1997).

Thus our research needs to incorporate the uncertainty associated with the weighting procedures used in survey analyses in the estimation of quantities of interest. Recall that in these research applications, the population that we wish to sample from is uncertain (typically a projection of likely voters in a subsequent election) and that in our case we do

not even know the population parameters with certainty (as we are using the CPS data to deal with the fact that the Hispanic population is rapidly changing in the United States), but that the weights developed and used in analyses are typically treated deterministically.

# 7 Conclusions and Future Directions

As we discussed earlier in this paper, appropriate weighting procedures can help improve the accuracy of inferences, when population parameters are known, when we have estimates of the same sample parameters, and when those demographic attributes we are weighting with are correlated with the behavioral quantity of interest. But often researchers are in a different situation, and they seek to weight to attributes that are not known; for example, in many situations (especially applied applications) researchers may wish to weight by attributes like partisanship to improve their inferences of behavioral quantities like voter candidate preferences.

Typically, such weighting methods are frowned upon, as the researcher rarely can assume that she has known measures of the partisan distribution in the population, either because it is not easily or accurately measured, or because it is a behavioral attribute as well, one that fluctuates over time. In these situations, weighting partisanship is frowned upon because it is difficult to ascertain the extent to which basic uncertainty about the population estimate of partisanship might lead behavioral inferences astray.

We believe developing ways in future research to take weighting uncertainty into account when weighting by demographic attributes, that those same methods might apply to this other situation, where attributes like partisanship are the subject of the weighting method. This would further improve our ability to weight samples to determine population values.

While our focus in this paper has been on survey samples of Hispanic voters, and the development of procedures to weight samples of respondents from that population, our methodologies and future research strategies have general applicability. Of course, one obvious next step will be for us to examine other samples of low-incidence populations; for example, voters from other low-incidence racial groups (Blacks and Asian-Americans), or other ethnic or demographic populations (Jewish, Muslim, or youth voters). All of these low-incidence populations are ones that have received little attention in the academic literature, but which are substantively and politically important populations whose attitudes and behaviors we need to better understand.

Furthermore, our methodologies should also apply to general population samples; like the National Election Study or the General Social Surveys. Can we improve the accuracy of the sample weights in those studies — and can we produce estimates of the various dimensions of uncertainty in those sample weights? While clearly outside the scope of our present study, these questions are also natural extensions of the research agenda we have presented in this paper.

# Notes

1. In this paper we use the term "Hispanic" to refer to persons or voters of Hispanic or Latino national origin or descent.

2. By comparison, 62% of the white citizen voting-aged population is estimated to have voted in 2000, and over 86% of the white registered citizen voting-aged population is thought to have voted in 2000. See U.S. Census Bureau (2002).

3. A secondary approach using RDD could involve what we will call "RDD Density Sampling", where telephone exchanges are sorted according to their density of Hispanic population, and only exchanges meeting some minimum population density requirement are included in the sample. A third approach based on RDD methodology is "RDD Disproportionate Stratified Sampling", where telephone exchanges are stratified based on their Hispanic population density, a sample is drawn from each strata, and the stratum are weighted to the correct population proportion. These latter two RDD approaches may also be expensive to implement, as they will still rely upon some filtering to get the sample to the appropriate criteria (Hispanic voters).

4. The sampling of low-incidence populations has received little attention in the social science survey methodology literature. Some attention has been paid to sampling of low-incidence populations using surname sampling designs (Himmelfarb, Loar and Mott, 1983; Shin and Yu 1984). In other fields, like epidemiology and health research, some attention has been directed to the question of surveying relatively rare populations (e.g., Kalsbeek 2003). Our paper is aimed at examination of the various alternatives for social scientific sampling of low-incidence populations, in this case, Hispanic voters in the United States.

5. As stated by Bush campaign advisor Matthew Dowd, July 17, 2004, on the Tim Russert Show.

6. For example, the 2004 National Election Survey post-election sample was weighted to age and education: "The 1,066 Post-Election cases were post-stratified to 2004 CPS March Supplement proportions for six (6) ages by four (4) education categories. The post-stratification compensates for differential non-response by age group and education level. The panel attrition weight for the Post-Election Study is the product of the Pre-Election final weight and the post- stratification factor formed by dividing the CPS proportion by the weighted NES proportion for each of the 24 age by education cells. The weight is scaled to sum to the number of cases, 1066." For further discussion of NES weighting procedures, see SRC 1998.

7. The scheme obviously generalizes straightforwardly to more dimensions, and the analyses we do in the paper are based on this scheme extended to four dimensions.

8. The weighted columns should obviously be the same for each survey: that is the point, to weight each survey to what we believe is truth. However, the minor differences across the weighted columns of the table arise from 'empty cells' in the survey data.

9. Details of this survey and the survey questionnaire can be found at http://www.democracycorps.com/reports/surveys/Democracy_Corps_Hispanic_Survey.pdf.

10. For additional details about this survey and the questionnaire, see http://www.democracycorps.com/reports/surveys/July_Hispanic_Survey.pdf.

11. For details on the NEP National exit poll methodology, see http://www.exit-poll.net/election-night/MethodsStatementNationalFinal.pdf.

12. The weights supplied with the NEP data are somewhat ambiguous in their origins, the result of a variety of corrections. In the NEP post-election analysis of their exit poll methodology, they state that the weighting process: "takes into account the probabilities of selection of the precinct and the sample voters within each sample precinct, the age-race-sex adjustment for non-interviews, the best estimate of the candidate vote percentages from each geographic region, and if applicable the portion of the vote that is being cast by absentee/early voters" (Edison Media Research and Mitofsky International 2005, page 9). This report also discusses significant problems in their weighting techniques regarding gender and absentee balloting (see page 5 for a summary discussion of these issues).

# 8 References


Alvarez, R. Michael and Lisa Garcia Bedolla. 2003. "The Foundations of Latino Voter Partisanship: Evidence From The 2000 Elections." *Journal of Politics* 65, 31-49.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41, 2, 641-674.

U.S. Census Bureau, 2002. "Voting and Registration in the Election of November 2000." Current Population Reports, Amie Jamieson, Hyon B. Shin and Jennifer Day, P20-542. Washington, D.C.: U.S. Department of Commerce.

Edison Media Research and Mitofsky International, 2005. "Evaluation of Edison/Mitofsky Election System 2004." http://www.exit-poll.net/election-night/EvaluationJan192005.pdf.

Himmelfarb, Harold S., R. Michael Loar, Susan H. Mott. 1983. "Sampling by Ethnic Surnames: The Case of American Jews." *The Public Opinion Quarterly*, 47, 2, 247-260.

Kalsbeek, William D. 2003. "Sampling Minority Groups in Health Surveys." *Statistics in Medicine* 22, 1527-1549.

Leal, David L., Matt A. Barreto, Jongho Lee, and Rodolfo O. de la Garza. 2005. "The Latino Vote in the 2004 Election." *PS: Political Science and Politics*, 38, 1, 41-50.

Levy, Paul S. and Stanley Lemeshow, 1999. *Sampling of Populations: Methods and Applications, Third Edition*. New York: John Wiley and Sons, Inc.

Mitofsky, Warren, Joel Bloom, Joseph Lenski, Scott Dingman, and Jennifer Agiesta, 2005. "A Test of Combined RDD/Registration-Based Sampling Model in Oregon's 2004 National Election Pool Survey: Lessons From A Dual Frame RBS/RDD Sample." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, May 12-15, 2005, Miami Beach, Florida.

Shin, Eui-Hang, Eui-Young Yu, 1984. "Use of Surnames in Ethnic Research: The Case of Kims in the Korean-American Population." *Demography* 21,3, 347-360.

Survey Research Center Staff, 1998. "Post-Stratified Cross-Sectional Analysis Weights for the 1992, 1994 and 1996 NES Data." NES Technical Report Series, No. nes010174.

Table 1: **Logit Estimates: Models of Turnout, November 2004 CPS**

| | Model based on: | |
|---|---|---|
| | **All-Voters** | **Hispanic Voters** |
| **age18to29** | -1.577** | -1.668** |
| | -51.8 | -14.41 |
| **age30to39** | -1.192** | -1.169** |
| | -39.06 | -9.99 |
| **age40to49** | -0.805** | -0.912** |
| | -26.64 | -7.6 |
| **age50to59** | -0.500** | -0.527** |
| | -15.05 | -3.95 |
| **lt hsgrad** | -2.519** | -2.229** |
| | -48.11 | -9.54 |
| **hsgrad** | -1.710** | -1.585** |
| | -35.54 | -6.87 |
| **somecoll** | -1.010** | -0.875** |
| | -20.65 | -3.74 |
| **collgrad** | -0.289** | -0.349 |
| | -5.43 | -1.36 |
| **married** | 0.535** | 0.428** |
| | -27.58 | -6.09 |
| **woman** | 0.144** | 0.093 |
| | -7.79 | -1.37 |
| **regdead** | -0.008** | -0.001 |
| | -6.91 | -0.17 |
| **Hispanic** | -0.272** | – |
| | -7.59 | – |
| **Constant** | 2.052** | 1.719** |
| | -27.92 | -4.74 |
| **Observations** | 64183 | 4122 |
| **PCF of Hispanic Voters** | 65.1% | 65.1% |

Table 2: **Weighted Versus Unweighted Sample Distribution**

| | **DC July 2004** | | **DC Feb 2004** | | **Exit 2004** | |
|---|---|---|---|---|---|---|
| | **Weighting Scheme** | | | | | |
| | **None** | **HV** | **None** | **HV** | **None** | **HV** |
| **Age:** | | | | | | |
| 18-29 | 13.1 | 20.2 | 10.9 | 20.8 | 19.4 | 20.8 |
| 30-39 | 17.1 | 22.0 | 17.7 | 21.8 | 18.8 | 22.0 |
| 40-49 | 21.1 | 22.2 | 20.8 | 21.9 | 23.8 | 21.8 |
| 50-59 | 20.7 | 16.2 | 22.9 | 16.4 | 19.2 | 16.3 |
| 60-98 | 28.0 | 19.4 | 27.7 | 19.1 | 18.9 | 19.1 |
| **Education** | | | | | | |
| LT HSgrad | 11.5 | 19.2 | 21.4 | 20.2 | 3.9 | 20.4 |
| HS Grad | 27.9 | 29.0 | 25.3 | 28.7 | 20.9 | 28.6 |
| Some Coll | 38.0 | 31.8 | 23.2 | 31.4 | 32.3 | 31.3 |
| Coll Grad | 24.6 | 14.0 | 23.0 | 13.9 | 26.3 | 13.8 |
| Post-Grad | 8.0 | 6.0 | 7.1 | 5.8 | 16.5 | 5.9 |

Cell entries are column percentages for the row variable in each of the three surveys.

Columns marked 'HV' are weighted to the model based only on Hispanic Voters. Other columns are unweighted.

Table 3: **Weighted and Unweighted Marginals - DC February 2004**

| | **Un Weighted** | **DCorps Weights** | **Hispanic Turnout Weights** | **All Voter Turnout Weights** |
|---|---|---|---|---|
| Dem | 58.2 | 59.5 | 58.1 | 58.0 |
| Ind | 13.8 | 15.7 | 14.8 | 14.7 |
| Rep | 25.5 | 22.0 | 24.8 | 24.9 |
| Other | 2.5 | 2.8 | 2.3 | 2.3 |
| Kerry-2 | 62.2 | 63.5 | 62.1 | 62.0 |

Entries are marginals computed with different weights.

Table 4: **Weighted and Unweighted Marginals - Democracy Corps July 2004**

| | **Un Weighted** | **DCorps Weights** | **Hispanic Turnout Weights** | **All Voter Turnout Weights** | **CPS Nov RAW Turnout Weights** |
|---|---|---|---|---|---|
| Dem | 62.5 | 62.8 | 63.4 | 63.3 | 63.6 |
| Ind | 13.1 | 13.1 | 14.2 | 14.2 | 14.6 |
| Rep | 24.3 | 24.1 | 22.4 | 22.5 | 21.7 |
| | | | | | |
| Kerry-2 | 66.8 | 67.6 | 68.7 | 68.5 | 69.5 |

Entries are marginals computed with different weights.

Table 5: **Weighted and Unweighted Marginals - VNS Exit Poll 2004**

| | **Un Weighted** | **VNS Weights** | **Hispanic Turnout Weights** | **All Voter Turnout Weights** | **CPS Nov RAW Turnout Weights** |
|---|---|---|---|---|---|
| Dem | 45.9 | 42.0 | 44.9 | 44.8 | 45.5 |
| Ind | 20.2 | 19.5 | 19.1 | 19.2 | 18.6 |
| Rep | 27.0 | 31.3 | 26.6 | 26.6 | 26.5 |
| Other | 7.0 | 7.2 | 9.5 | 9.4 | 9.4 |
| Kerry-2 | 61.5 | 53.5 | 59.8 | 60.0 | 59.4 |

Entries are marginals computed with different weights.