YouGovPolimetrix

Sampling and Weighting Methodology
for the October 2009 Texas Statewide Study

Survey Data and Methodology

The October 2009 Texas Statewide Study was designed by researchers in the UT-Austin Department of Government and conducted by YouGov/Polimetrix, a firm with demonstrated success in internet polling. YouGov/Polimetrix accomplishes internet polling through a unique sampling procedure known as "matched random sampling." The firm begins with two lists: (1) a list of all adult "consumers" in Texas (covering approximately 95 percent of the adult population), and (2) a list of people who have agreed to take YouGov/Polimetrix's surveys. For each list, Polimetrix has an extensive set of demographics.

The sampling procedure then progresses in two stages. First, a random sample of consumers is drawn. For each person drawn from this sample a list of key demographics is recorded. In essence, each individual drawn is represented as a cluster of demographic characteristics, including age, income, education, race, gender, longitude and latitude, etc. Second, YouGov/Polimetrix uses a matching algorithm to find the PollingPoint panelist who is the closest match to the person drawn off the consumer file. In this way an entire "matched" random sample is constructed for all people in the "drawn" sample.

The October 2009 poll consists of 800 registered voters from the state of Texas, and has a margin of error of +/- 3.46 percentage points at the 95% confidence level. For the Republican Primary items, 359 registered voters said they were “definitely” or “probably” going to vote in the March 2010 Republican Primary (the attendant margin of error is +/- 5.17 points). For the Democratic Primary items, 247 registered voters said they were “definitely” or “probably” going to vote in the March 2010 Democratic Primary (the attendant margin of error is +/- 6.24 points). Response rates are almost 100% given the matching methodology. The YouGov/Polimetrix pool includes people who are much less likely to have access to the Internet or a personal computer. YouGov/Polimetrix has been especially assiduous at enlisting ethnic and racial minorities, as well as people who are less affluent, as part of their attempt to ensure the representativeness of their samples. Surveys were completed between October 20 and October 27, 2009.

Polimetrix interviewed 1,152 respondents who were then matched down to a sample of 800 to produce the final data set. The respondents were matched on gender, age, race, education, party identification and political interest. YouGov/Polimetrix then weighted the matched set of survey respondents to known marginals for the registered voters of Texas from the 2008 Current Population Study. Those marginals are:

Age:

18-34: 27.0%

35-54: 38.3%

55+: 34.7%

Gender:

Male: 46.4%

Female: 53.6%

Race:

White/Other: 66.2%

Black: 13.8%

Hispanic: 20.0%

Education:

HS or less: 37.2%

Some College: 33.6%

College Graduate: 20.9%

Post-graduate: 8.2%

Survey Panel Data

The PollingPoint panel, a proprietary opt-in survey panel, is comprised of 1.6 million U.S. residents who have agreed to participate in YouGov/Polimetrix's Web surveys. At any given time, YouGov/Polimetrix maintains a minimum of five recruitment campaigns based on salient current events.

Panel members are recruited by a number of methods and on a variety of topics to help ensure diversity in the panel population. Recruiting methods include Web advertising campaigns (public surveys), permission-based email campaigns, partner sponsored solicitations, telephone-to-Web recruitment (RDD based sampling), and mail-to-Web recruitment (Voter Registration Based Sampling).

The primary method of recruitment for the PollingPoint Panel is Web advertising campaigns that appear based on keyword searches. In practice, a search in Google may prompt an active PollingPoint advertisement soliciting opinion on the search topic. At the conclusion of the short survey respondents are invited to join the PollingPoint panel in order to receive and participate in additional surveys. After a double opt-in procedure, where respondents must confirm their consent by responding to an email, the database checks to ensure the newly recruited panelist is in fact new and that the address information provided is valid.

Additionally, YouGov/Polimetrix augments their panel with difficult to recruit respondents by soliciting panelists in telephone and mail surveys. For example, in the fall and winter of 2006, YouGov/Polimetrix completed telephone interviews using RDD sampling and invited respondents to join the online panel. Respondents provided a working email where they could confirm their consent and request to receive online survey invitations. YouGov/Polimetrix also employed registration based sampling, inviting respondents to complete a pre-election survey online. At the conclusion of that survey, respondents were invited to become PollingPoint members and receive additional survey invitations at their email address.

The PollingPoint panel currently has over 55,000 active panelists who are registered voters in Texas. These panelists cover a wide range of demographic characteristics.

Sampling and Sample Matching

Sample matching is a methodology for selection of "representative" samples from non-randomly selected pools of respondents. It is ideally suited for Web access panels, but could also be used for other types of surveys, such as phone surveys. Sample matching starts with an enumeration of the *target population.* For general population studies, the target population is all adults, and can be enumerated through the use of the decennial Census or a high quality survey, such as the American Community Survey. In other contexts, this is known as the *sampling frame*, though, unlike conventional sampling, the sample is *not* drawn from the frame. Traditional sampling, then, selects individuals from the sampling frame at random for participation in the study. This may not be feasible or economical as the contact information, especially email addresses, is not available for all individuals in the frame and refusals to participate increase the costs of sampling in this way.

Sample selection using the matching methodology is a two-stage process. First, a random sample is drawn from the target population. We call this sample the *target sample.* Details on how the target sample is drawn are provided below, but the essential idea is that this sample is a true probability sample and thus representative of the frame from which it was drawn.

Second, for each member of the target sample, we select one or more *matching* members from our pool of opt-in respondents. This is called the *matched sample.* Matching is accomplished using a large set of variables that are available in consumer and voter databases for both the target population and the opt-in panel.

The purpose of matching is to find an available respondent who is as similar as possible to the selected member of the target sample. The result is a sample of respondents who have the same measured characteristics as the target sample. Under certain conditions, described below, the matched sample will have similar properties to a true random sample. That is, the matched sample mimics the characteristics of the target sample. It is, as far as we can tell, "representative" of the target population (because it is similar to the target sample).

When choosing the matched sample, it is necessary to find the closest matching respondent in the panel of opt-ins to each member of the target sample. Polimetrix employs the proximity matching method to find the closest matching respondent. For each variable used for matching, we define a *distance function*, d(x,y), which describes how "close" the values x and y are on a particular attribute. The overall distance between a member of the target sample and a member of the panel is a weighted sum of the individual distance functions on each attribute. The weights can be adjusted for each study based upon which variables are thought to be important for that study, though, for the most part, we have not found the matching procedure to be sensitive to small adjustments of the weights. A large weight, on the other hand, forces the algorithm toward an exact match on that dimension.

Sampling Frame and Target Sample

YouGov/Polimetrix constructed a national sampling frame from the 2006 American Community Survey, including data on age, race, gender, education, marital status, number of children under 18, family income, employment status, citizenship, state, and metropolitan area. The frame was constructed by stratified sampling from the full 2006 ACS sample with selection within strata by weighted sampling with replacements (using the person weights on the public use file). Data on voter registration status and turnout were matched to this frame using the November 2008 Current Population Survey. Data on interest in politics and party identification were then matched to this frame from the 2004 National Annenberg Election Study, using the following variables for the match: age, race, gender, education, marital status, number of children under 18, family income, employment status, citizenship, state. The target sample of 800 Texas registered voters was selected with stratification by age, race, gender, education, and with simple random sampling within strata.

Weighting

Because matching is approximate, rather than exact, and response rates vary by group, the sample of completed interviews normally shows small amounts of imbalance that can be corrected by post-stratification weighting.

Raking, first proposed by Deming and Stephan (1940), adjusts an initial set of weights to match a known set of population marginals, using a method of iterative proportional fitting (see Bishop, Fienberg and Holland, 1975 for details). In this procedure, the weights are adjusted sequentially to match the marginal distribution of each weight variable. The process proceeds until all marginals are matched. It does not require any information about the joint distribution of the variables (though, if these data are available and believed to be important, they can be employed by defining a marginal distribution involving a cross-classification of two variables).

We calculated post-stratification weights by raking the completed interviews to known marginals for the general population of Texas from the November 2008 Current Population Survey for the following variables: age, race, gender, and education.