Inference From Matched Samples in the 2008 U.S. National Elections*

Douglas Rivers† Delia Bailey‡

Abstract

The performance of matched sampling is assessed using data from the 2008 U.S. Presidential election. The assumptions necessary for the validity of matched sampling, including ignorability, are described. With a matching ratio of about five, the matched sample reproduces the joint demographic distribution of the target population very closely. The sampling distribution of the associated state-level vote estimates is approximately standard normal with unit variance, suggesting little or no selection bias conditional upon a full set of demographic controls. The results are compared to RDD telephone and Internet samples, none of which are clearly better and some (such as the 2008 ANES Internet panel) are substantially worse.

Key Words: sampling, matching, propensity score, surveys

1. Introduction

2008 was the year that Internet polling came of age. Most presidential campaigns conducted at least some polling on the Internet. *The Economist*, the Associated Press, and CBS News conducted Internet surveys throughout the campaign. Several large academic projects, including the American National Election Study (ANES), the National Annenberg Election Study (NAES), the Cooperative Congressional Election Study (CCES), and the Cooperative Campaign Analysis Project (CCAP) all collected data using the Internet.

Unlike telephone polling, where similar sampling and weighting methods are used by most survey organizations, there is a wide discrepancy between how Internet surveys are conducted. Knowledge Networks (KN) uses a traditional sampling methodology—random digit dial (RDD)—and provides Internet access to selected respondents who do not currently have it. This is comfortably within the mainstream of survey sampling practice, as the sampling frame includes households without Internet access and uses probabilistic selection.

The KN approach has been shown to produce similar results to RDD telephone surveys, but it also shares most of the problems of telephone interviewing (such as low response rates and relatively high costs) as well as the usual problems of access panels (attrition and geographical limitations). Matched samples allow low cost opt-in samples to be used for both descriptive population estimates and analytic studies. The purpose of this paper is to assess the effectiveness of matching and weighting to remove selection biases in polling applications.

Because of low recruitment response rates and attrition, it is very difficult to operate an access panel with anything approximating "known" probabilities of selection. Even very expensive RDD panels (such as that built by KN for ANES in

*The views expressed here are those of the authors, not of any organizations they are currently or formerly associated with.

†Department of Political Science, Stanford University, Stanford, CA 94305, and YouGov Polimetrix, 285 Hamilton Ave., Palo Alto, CA 94301.

‡YouGov Polimetrix, 285 Hamilton Ave., Palo Alto, CA 94301.

2008) do not produce samples that are demographically representative. If subsamples were selected at random from the panel, they would consistently reflect the biases that have crept into the panel. Consequently, the selection probabilities are "balanced" to reflect known population characteristics. The quality of the resulting estimates depend on the quality of the modeling, not the selection probabilities.

The estimates from YouGov Polimetrix are based instead on a sample matching methodology. A synthetic sampling frame is constructed starting with the 2005–2007 American Community Study, a high quality probability sample. Variables from other sources, including the 2004 (and now 2008) CPS Registration and Voting Supplement, the Pew Study of Religious Life, and state voter lists, have been matched to it (using nearest neighbors matching). We operate a large opt-in panel (with over one million members), recruited from a diverse array of sources, so that it has sufficient numbers in nearly any category determined by the cross-classification of the frame variables. To draw a sample of size n from the panel, we first draw a (stratified) sample of the same size from the frame.[1] Then we use Mahalanobis distance matching to choose the closest n units from the panel. The matches are imperfect, but most discrepancies can be addressed by propensity score weighting.

The sample matching method differs from both the ANES Internet Panel in that it does not use random selection at its initial stage of recruitment. Instead, a large pool of respondents (over one million) is recruited through Internet advertising and then a subsample is selected by matching to a random sample drawn from the target population. This is a form of purposive selection intended to match the joint distribution of a set of covariates in the target population. Under an assumption of ignorability and some technical conditions (discussed below), the matching estimator is consistent, asymptotically normally distributed, with a covariance matrix that can be estimated robustly.

This approach is controversial because the method of selection is purposive. The advantage of the matching method is that it is feasible to balance a non-random sample on a large number of covariates to reduce potential bias. The resulting samples are much less expensive to collect and, in terms of covariate balance, are superior to what is currently achievable with RDD. The disadvantage, of course, is that the validity of the estimates and associated inferences depend upon the validity of the matching and weighting models which are difficult to test. However, all methods with substantial non-response, including all RDD studies conducted today, depend heavily upon modeling assumptions.

In evaluating alternative methodologies for Internet polling, we shall focus upon their statistical properties. Nonresponse and self-selection are analyzed in a model-based framework. The key assumption supporting this analysis is *ignorability* of selection conditional upon a set of covariates. The validity of this assumption is tested using different sets of covariates. However, ignorability does not hold exactly and some estimates of the magnitude of biases will be obtained.

2. Data

Cooperative Congressional Election Study. The 2008 Cooperative Congressional Election Study (CCES) was conducted by YouGov for a consortium of 30 universities. The Principal Investigator was Stephen Ansolabehere of M.I.T. and followed a similar design to the 2006 CCES (Vavreck and Rivers, 2008) with a pre-election

[1] For election studies, we typically stratify on race × education × registration × gender × age, and collapse the tree from the bottom so that each cell has a minimum of ten units.

interview of approximately 20 minutes and a post-election interview of approximately 10 minutes. The study was conducted on YouGov's PollingPoint panel and the data reported here are from the second release of the dataset, containing 34,800 interviews.

The YouGov PollingPoint panel is an "opt-in" panel recruited primarily using Internet advertising. The relevant feature is not that panelists "opt-in" (all panels of humans are "opt-in"), but that the initial stage of selection lacks a well-defined sampling frame, which makes it impossible to calculate a meaningful response rate for panel recruitment. Recruitment occurs on a continuous basis and the set of 1.5 million panelists are used as a pool from which respondents can be selected for individual studies.

Panelists are selected for individual studies using the matching methodology described in Rivers (2007). A target population is selected by drawing a stratified random sample from the 2005–2007 American Community Survey (ACS) public use microfile. Voter registration was imputed onto the file by matching the sample with the November 2008 Current Population Survey (CPS) Voting and Registration Supplement.[2] News interest, third party identification, ideological non-placement (respondents who are unable to place themselves on a five-point liberal–conservative scale), church attendance, and born-again/evangelical status were imputed by matching to the 2007 Pew Religious Life survey. A total of 34,800 respondents were selected from the panel by matching on these variables. The sample was stratified on registration and race. The matched sample was weighted using propensity scores on the same set of variables.

Matched Sample. For the purpose of comparison with the other pre-election surveys, a matched sample of 3,000 respondents was selected from the pool of 34,800 respondents in the CCES dataset. First a sample is drawn from the frame, stratified on geographic region, gender, race, education, and age. Then, a matched sample is selected from the CCES pool, by nearest neighbor matching using the Mahalanobis distance metric on age, race, gender, education, marital status, employment status, geographic region, state, level of news interest, born again/evangelical status, and ideological non-placement. An estimated propensity score is used to produce weights for the matched sample, utilizing the same variables as in the matching distance measure.

American National Election Studies Internet Panel. The 2008–09 American National Election Studies (ANES) Internet Panel was conducted by Knowledge Networks (KN) under a grant from the National Science Foundation. This panel should be distinguished from the KN access panel, which was used for the National Annenberg Election Study (NAES) and the Associated Press–Yahoo! poll. 2,371 panelists were recruited by random digit dialing in 2007 and an additional 1,850 panelists were recruited in Summer 2008. The panelists completed an online profile survey in January 2008, for the first cohort, and in September 2008, for the second cohort.

The sample weights consist of a base weight that adjusts for differences in probability of selection, and post stratification weights produced by raking the base weights to target marginals from the Current Population Survey on gender, census region, age, race and ethnicity, and educational attainment. The final weights were trimmed at five.

USA Today–Gallup Final Pre-election Poll. Gallup and *USA Today* conducted

[2]The pre-election results, released on November 1, 2008, used registration data from the 2004 CPS Voting and Registration Supplement.

their final pre-election poll between October 31 and November 2, 2008. The sample was selected by random digit dialing from numbers stratified by landline and cellphone exchanges. Quota sampling (by gender) was used at the final stage for within-household selection. A total of 3,050 interviews were completed.

CNN Final Pre-election Poll. CNN conducted their final pre-election poll between October 30 and November 1, 2008 through the Opinion Research Corporation. The sample was selected by random digit dialing, selecting the youngest respondent in the household, with selection on gender. A total of 1,017 interviews were completed.

NBC/Wall Street Journal Final Pre-election Poll. NBC and the *Wall Street Journal* conducted their final pre-election poll on November 1 and 2, 2008. The sample was selected by random digit dialing numbers stratified by region and place size, with an additional sample drawn from a list of cell phone users who were screened to select cell-phone only users. At the household level, registered voters who were at least 18 years of age were selected on gender. A total of 1,011 likely voters were interviewed, including 108 cell phone only users. The post stratification weights were constructed to balance on gender, age, race, ethnicity, education, and evangelicals.

ANES 2008 Time Series Study. The American National Election Studies (ANES) conducted the pre-election portion of the panel survey between September 2 and November 3, 2008. The ANES conducted 2,323 in-person interviews in the pre-election wave. The sample utilizes a complex design and contained an oversample of African American and Latino adults. The response rate was 60%.

Current Population Survey November 2008 Voting and Registration Supplement. The Current Population Survey, conducted by the Bureau of the Census, is also a complex design, with a sample of 56,000 households. The Voting and Registration survey adds data on reported voter registration and turnout. The interviews were conducted between November 16 and 22, 2008. The response rate is 93% and the turnout estimates are consistent with tabulated vote counts.

National Election Pool Exit Poll. The National Election Pool (NEP) conducted an exit poll (and, in 18 states, a telephone survey of early voters). The survey utilizes a complex design with geographic stratification within states and probability proportional to size selection of clusters (precincts) within strata. The data are raked to interviewer counts of refusals and misses according to age, race, gender, and vote at the stratum level. A public use microdata file has not yet been released. The data reported here are from 17,836 respondents from a subscriber web site. The exit polls have a fairly high response rate (44% in 2008) and do not rely upon self-report for turnout (except for early voters).

3. Sample Balance

All of the surveys that we discuss come with a set of weights created by the data producer. These weights are generally *not* the reciprocal of the selection probabilities. In some cases, the weight may reflect a post-stratification adjustment to a base weight, which was an inverse probability weight. However, in many cases the probability of selection has not been used at all.[3] The purpose of weights is primarily to adjust for nonresponse. For matched samples, the purpose of weights is to remove bias due to partial or imperfect matching.

[3] As evidenced by all respondents having identical demographics also having identical weights.

The goal of sample matching is to provide a sample that is representative of the joint distribution of the covariates in the population. With a large and diverse enough pool, it is feasible to match the joint distribution of a large set of covariates closely. With probability samples, representation is obtained by using stratification and, within strata, random selection. In fact, all of the samples other than CPS suffer from substantial nonresponse, so that random selection provides no guarantee of representativeness.[4] RDD designs usually do not involve any stratification on demographics, so it is conventional to use some form of post-stratification to correct for nonresponse.

In this section, we investigate the effectiveness of matched sampling in reproducing population demographics. The demographic composition of matched samples is compared with both area probability samples (CPS and ANES) and RDD-based samples (the pre-election media telephone polls as well as the ANES Internet panel). We shall attempt to distinguish between representativeness obtained from random selection and by post-stratification.

We will use the CPS November 2008 Voting and Registration Supplement as a baseline for comparison. This survey has a high response rate (about 93%) and, after weighting[5], produces estimates close to the actual vote totals. Most of the other surveys appear to have been weighted to CPS marginals on some variables.

Ideally, the unweighted matched sample would be compared to the probability samples weighted by the reciprocal of their selection probabilities. Unfortunately, of the studies analyzed here, only the ANES Internet panel provides selection probabilities. The selection probabilities for the RDD samples should be proportional, in theory, to the ratio of the number of telephone lines to the number of eligible adults in the selected households. In practice, because of a correlation between household size and contact rates, the actual selection probability is weakly related to this ratio and it is often ignored. Some of the phone surveys also have a cell phone (or cell phone only) stratum, but the stratum allocations are intended to be self-representing and can be ignored.

The ANES face-to-face survey (which we refer to as "ANES" without any additional qualification) includes an oversample of African Americans and Latinos, but the current version of the file and documentation do not provide enough information to compute the correct selection probabilities within stratum. We have post-stratified the entire sample (including non-voters) using the November 2008 CPS race distribution. We shall refer to this as "unweighted," though it is actually weighted by race.

The target populations of the surveys described in Section 2 vary between adults, adult citizens, registered voters, and likely voters. For consistency, we will limit comparisons to voters (or likely voters for the pre-election polls). The top panel of Table 1 shows the unweighted distribution of voters (or likely voters) in each sample for three racial groups and five education levels.

The unweighted RDD samples consistently under-represent minorities and low education respondents. The size of the discrepancy is quite large, especially among persons without a high school degree and Hispanics. For example, the ANES Internet panel under-represents the lowest education category by 71%, high school graduates by 56%, Hispanics by 45%, and Blacks by 53%. This occurs despite what

[4]Cochran (1973) shows that model-free inferences are impractical if nonresponse exceeds about 10%. Even the ANES area probability sample does not meet this standard.

[5]CPS uses a complex design and cannot be analyzed without use of the stratum and selection weights.

Table 1: Sample Composition: Voters

	Race			Education				
Unweighted	White	Black	Hispanic	<HS	HS	Some Col.	College	Postgrad
ANES	79.6%	12.5%	7.9%	7.0%	29.3%	34.0%	20.3%	9.3%
ANES Internet	90.2%	5.7%	4.1%	2.0%	13.9%	37.0%	26.3%	20.8%
CNN	89.7%	7.0%	3.2%	2.4%	21.3%	29.2%	25.6%	21.5%
CPS	80.5%	12.1%	7.4%	6.9%	27.4%	31.6%	22.4%	11.7%
Gallup	87.8%	8.4%	3.8%	2.5%	16.6%	31.2%	27.0%	22.7%
NBC/WSJ	75.8%	14.8%	9.5%	4.0%	22.9%	26.0%	27.0%	20.2%
Pew	85.6%	9.7%	4.8%	4.5%	23.7%	29.1%	25.6%	17.2%
Matched	81.3%	11.6%	7.1%	6.5%	31.2%	33.1%	19.1%	10.2%
Weighted								
ANES	81.0%	12.3%	6.7%	7.8%	28.5%	30.8%	22.0%	11.0%
ANES Internet	82.4%	11.2%	6.4%	5.8%	28.8%	31.4%	22.5%	11.5%
CNN	78.1%	12.9%	9.0%	6.9%	28.0%	31.3%	18.0%	15.7%
CPS	80.5%	12.1%	7.4%	6.9%	27.3%	31.6%	22.4%	11.8%
Gallup	81.9%	12.0%	6.1%	6.8%	26.6%	30.9%	20.2%	15.5%
NBC/WSJ	80.0%	11.5%	8.5%	3.7%	24.0%	27.7%	26.6%	18.1%
Pew	80.0%	11.3%	8.8%	7.9%	29.3%	30.4%	19.8%	12.6%
Matched	80.5%	11.7%	7.7%	9.0%	28.4%	32.8%	19.4%	10.4%

is considered a respectable response rate (about 28%, assuming a 95% eligibility rate for non-contacts) by current standards.

The lower panel of Table 1 shows the weighted distribution of race and education in the samples. After weighting, the sample proportions of Black and Hispanic voters is close to the corresponding proportions in CPS. In some cases, however, the education distributions are off by a significant amount. This probably results from collapsing of the adjustment cells or from trimming the weights not to exceed some prescribed value.

Table 2 shows the distributions of gender, marital status, and region within the various samples. RDD samples that use proper within household selection procedures invariably over-represent women. Some of the phone samples resort to gender quotas (Gallup documents this practice) or thinly disguised equivalents (such as selecting men at "randomly" with higher probability than women and then ignoring the selection probability in estimation). All of the samples have been post-stratified on gender, so the discrepancies are nearly eliminated in the weighted distributions shown in the lower panel. The same is not true of marital status. Neither the of the ANES samples nor Gallup appear to have used marital status in their weighting algorithms.

Next, we turn to the age and income distributions in the various samples. These are continuous variables so it's feasible to compare the sample and population densities, as shown in Figure 1. In Figure 1 (and subsequent figures), a dark black line is used for CPS, a red line for the matched sample, and gray lines for the remaining surveys.

It is evident that all of the non-matched samples substantially under-represent younger voters and over-represent older voters. With one exception, the weighted distributions are much closer to the actual distributions, though all of the samples (including the matched sample) miss the "hump" around age fifty in the CPS distribution.

Table 2: Sample Composition: Voters

	Gender	**Marital Status**	**Region**			
	Male	Married	Northeast	Midwest	South	West
ANES	41.9%	48.4%	10.8%	19.2%	43.3%	26.9%
ANES Internet	44.1%	76.2%	16.0%	28.2%	31.8%	24.0%
CNN	49.0%	—	16.5%	22.2%	37.0%	24.2%
CPS	46.3%	60.5%	18.2%	23.9%	36.2%	21.7%
Gallup	50.0%	59.1%	20.7%	21.9%	34.2%	23.1%
NBC/WSJ	48.7%	—	18.6%	22.8%	35.7%	22.9%
Pew	47.2%	61.7%	17.2%	25.1%	37.1%	20.6%
Matched	47.3%	57.9%	18.3%	23.7%	37.1%	20.9%
Weighted						
ANES	42.9%	53.6%	13.5%	21.1%	43.0%	22.5%
ANES Internet	47.5%	72.4%	18.5%	24.4%	35.8%	21.4%
CNN	47.4%	—	17.7%	21.0%	38.3%	23.0%
CPS	46.3%	60.5%	18.2%	23.9%	36.2%	21.7%
Gallup	46.9%	55.0%	19.9%	22.1%	35.3%	22.8%
NBC/WSJ	48.0%	—	18.3%	25.0%	34.3%	22.4%
Pew	47.8%	57.4%	17.6%	24.1%	36.6%	21.6%
Matched	47.1%	56.7%	18.1%	23.7%	37.2%	21.0%

It is, of course, hardly news that RDD encounters difficulties representing age, race, gender, and education and standard weighting procedures have long been employed to address these problems. However, it is less well understood that the weighting methods (raking on age and gender marginals, for example) fail to reproduce the *joint* distribution of the variables very well. The RDD samples consistently over-estimate the proportion of young male voters and under-estimate the proportion of female voters. This could be fixed by raking on the cross-classification of gender and age or using propensity scores.

The matched sample requires little weighting (the maximum propensity score weight is about two, compared to weights that have been *trimmed* at levels two or three times higher in the RDD samples) and represents all categories shown well.

4. Ignorability

The bias of estimates from samples with nonresponse or self-selection depends upon the pattern of missingness. In Rubins terminology, data are missing completely at random (MCAR) if the both the covariates x and the survey measurements y are uncorrelated with the selection indicator r. Under the weaker condition of missing at random (MAR), where the survey measurements y are only assumed to be conditionally independent of the selection indicator r (conditional upon the covariates x), it is possible to remove the bias by weighting or subclassification upon the covariates or the propensity score. In either case, nonresponse or self-selection is said to be *ignorable*, in the sense that the analysis can be carried out using the likelihood for the complete data.

The ignorability condition can be formulated in several different, but equivalent, ways. For example, conditional independence of selection and the survey measurement y given the covariates x implies that

$$E(r|x, y) = E(r|x) \tag{1}$$

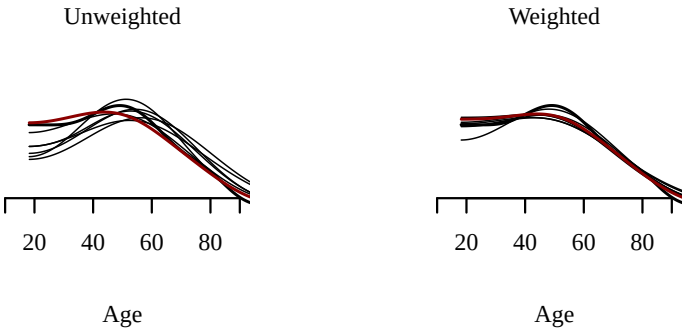

Figure 1: Age Distributions

Table 3: Logit Coefficients from Ignorability Analysis

	Constant	Black	Hispanic	Female	Age	Educ	N	log L
ANES	-0.25	4.90	1.24	0.20	-0.014	-0.04	1509	-893.37
	(0.09)	(0.81)	(0.24)	(0.11)	(0.003)	(0.05)		
ANES Internet	-0.43	4.07	0.86	0.37	0.0001	0.23	2013	-1217.95
	(0.07)	(0.44)	(0.20)	(0.10)	(0.003)	(0.04)		
CNN	-0.31	18.57	1.21	0.27	-0.001	0.21	678	-391.93
	(0.13)	(981.1)	(0.31)	(0.17)	(0.01)	(0.07)		
Gallup	-0.20	4.07	1.33	0.29	-0.005	0.13	2269	-1361.95
	(0.07)	(0.45)	(0.21)	(0.09)	(0.003)	(0.04)		
NBC/WSJ	-0.31	3.49	0.94	0.28	-0.008	0.20	924	-567.32
	(0.11)	(0.52)	(0.27)	(0.14)	(0.004)	(0.06)		
Pew	-0.26	3.36	0.85	0.33	-0.009	0.07	2323	-1435.72
	(0.07)	(0.33)	(0.16)	(0.09)	(0.003)	(0.04)		
Matched	-0.34	3.73	0.57	0.38	-0.02	0.14	2762	-1682.79
	(0.06)	(0.34)	(0.15)	(0.08)	(0.002)	(0.04)		

which could be tested using a choice-based sampling estimator (Manski and McFadden, 1981). A somewhat simpler method is just to test whether the conditional distribution of y given x is the same in each sample. We use a linear logistic regression of vote on the set of demographic covariates to test whether these conditional distributions are the same across samples. The results are shown in Table 3 below. The logits were weighted as described in the preceding section.

For the most part, the logistic regression results are similar across the different samples, suggesting that ignorability holds approximately. There is one serious discrepancy. Years of schooling has a small (and insignificant) negative effect on Obama vote in the ANES sample, while all of the other samples show a positive and usually significant coefficient for years of schooling. The ANES estimate is rather implausible and conflicts with the exit poll.

5. Empirical Sampling Distributions

Model-based estimates depend upon assumptions which can be difficult to check. We gain confidence in the usefulness and reliability of a model by assessing its

Table 4: Standardized Errors

	Bias		Standard Error		
Survey	Percent	SD	SRS	VIF	**RMSE**
ANES Internet	−4.9%	−0.46	1.47	1.10	0.10
Gallup	6.0%	0.28	1.47	1.22	0.10
Pew	3.9%	0.30	1.13	0.99	0.07
Matched	1.8%	0.14	0.96	0.94	0.06

performance in repeated application. We can compare state level estimates with the reported vote to obtain empirical estimates of the sampling distributions of the different procedures. In this section, we check whether these sampling distributions approximate the theoretical standard normal limiting distribution that would hold under an assumption of ignorable selection.

Because the surveys used different sample sizes and weighting methods, the survey errors have different sampling distributions. For each survey we have produced an estimate $\hat{\theta}$ for each state s and estimated its variance by $\hat{V}(\hat{\theta}) = (1+s_w^2)\hat{\theta}(1-\hat{\theta})$, where s_w is the standard deviation of the weights.[6] Since the observations are independent, the Central Limit Theorem implies that the errors are approximately normally distributed if the sample size is sufficiently large. Following the customary rule of thumb that the normal approximation is valid for thirty or more observations if the proportions are not too extreme, we restrict attention to states where more than thirty interviews were conducted for a particular survey.

In this section, we exclude the ANES area probability sample because it is clustered and impossible to calculate reliable standard error estimates for individual states. The other samples are not clustered, so reasonable standard error estimates can be calculated for states with more than 30 interviews. We omit the CNN and NBC/*Wall Street Journal* surveys because their small sample size means that too few state level estimates can be produced to obtain a reasonable density estimate for the sampling distribution. The results are shown in Figure 3.

The matched and Pew samples exhibit nearly symmetric, unimodal distributions, while the ANES Internet and Gallup samples are somewhat skewed (with ANES Internet having a negative skew and Gallup a positive skew). This is not too surprising, since both the matched sample and Pew accurately estimated the size of Obamas victory, while Gallup over-estimated and the ANES Internet panel underestimated the size of Obamas margin.

The matched sample produced by far the smallest percentage bias—only 0.3%—though there was also no significant bias for the Pew survey, as shown in the first two columns of Table 4.

We can also use this analysis to evaluate alternative standard error estimates. Most survey organization report a margin of error using the formula which assumes simple random sampling without weighting, which is incorrect, as shown by the column labeled "SRS" in Table 4. For the three RDD-based samples, the SRS formula underestimates actual sampling variation by between 14% and 48%. The standard error estimate used with the variance inflation factor adjustment (labelled

[6]The variance inflation factor $1+s_w^2$ assumes the weights have been normalized to sum to sample size. This formula assumes simple random sampling and should tend to overstate the variance of the sampling distribution under conventional post-stratification assumptions. See Gelman and Little (1998) and Little and Vartivarian (2005) for further discussion.

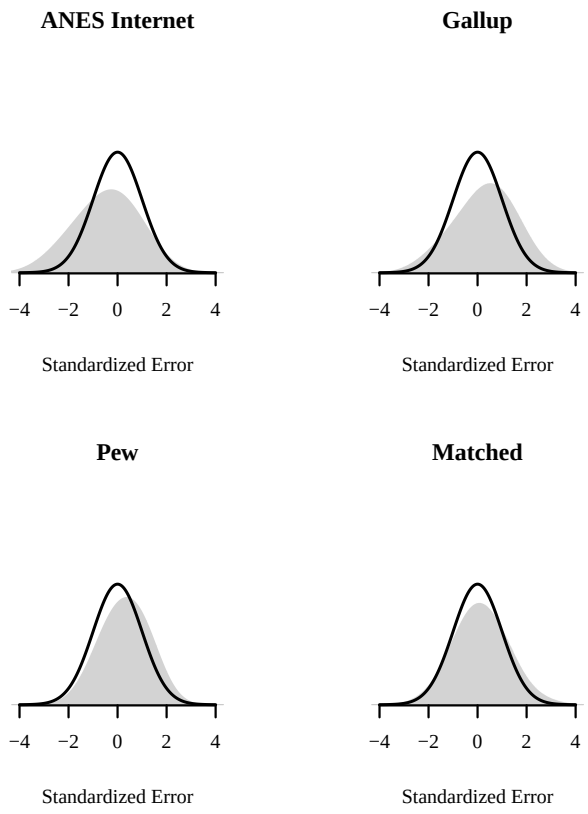

Figure 2: Sampling Distributions of Obama Vote Estimates

"VIF" in Table 4) is fairly close to one for all of the samples. This appears to give a serviceable estimate.

Finally, the last column of Table 4 displays the root mean square error (RMSE) of the state level estimates for the standardized error (the discrepancy between the poll estimate and the outcome divided by its estimated standard error). This combines both the (squared) bias and sampling error and provides an overall estimate of survey accuracy. The matched sample and Pew significantly outperform the other two surveys.

In the case of the matched sample and Pew, the standard normal limiting distribution appears to provide a sound basis for inference. The margins of error typically reported for RDD surveys (assuming SRS without weighting) grossly overstate the coverage of confidence intervals. The confidence intervals computed using the robust standard error (which have been used at YouGov Polimetrix in our press releases) have the stated level of coverage.

6. Domain Estimates

In this section we examine Obama vote estimates within various demographic groups (domains) in the surveys. The published exit poll estimates provide a plausible benchmark for comparison, because these have been post-stratified (at the stratum level) for refusals and misses and election returns. It is possible that there is non-ignorable selection causing bias in domain estimates, but which cancel out in state level estimates. This provides a further test of ignorability of selection, since the

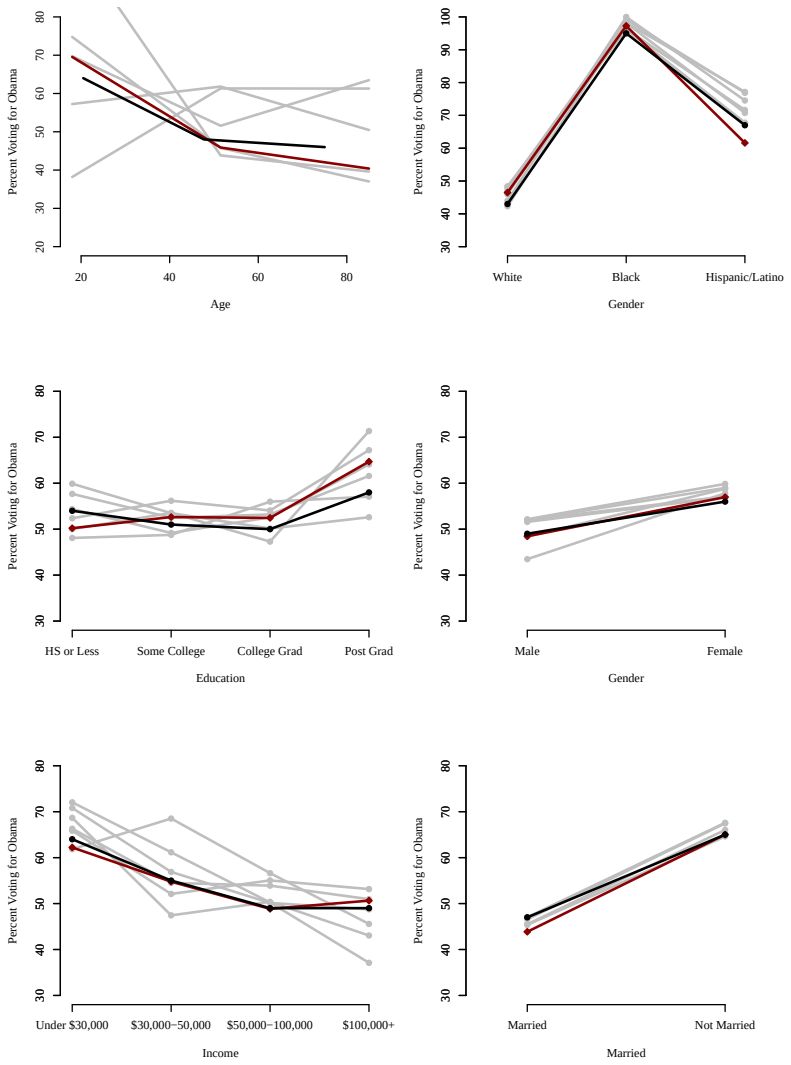

Figure 3: Percentage Voting for Obama

domain estimates should be close to the exit poll if ignorability holds.

The matched sample provides nearly identical estimates to the exit poll for gender, income, and marital status. The Obama vote estimate from the matched sample is very close to the exit poll for whites and blacks, but five percent lower for Hispanics. The matched sample estimates a larger difference between the Obama vote percentage for high and low education respondents than the exit poll, though the difference is not large.[7]

For each of the domains, the matched sample estimates is within the range of the estimates from the other surveys and often closer to the exit poll than the RDD samples. There is wide variation in the RDD estimates, some of which are substantially worse than the matched sample.

[7]High education respondents have much higher response rates than low education respondents in the exit poll. However, the exit poll is not weighted by education (since the interviewers are unable to estimate the education level of refusals and misses). The effect of post-stratifying on vote when there is a positive correlation with the confounding variable is to attenuate the relationship, which may explain the discrepancy.

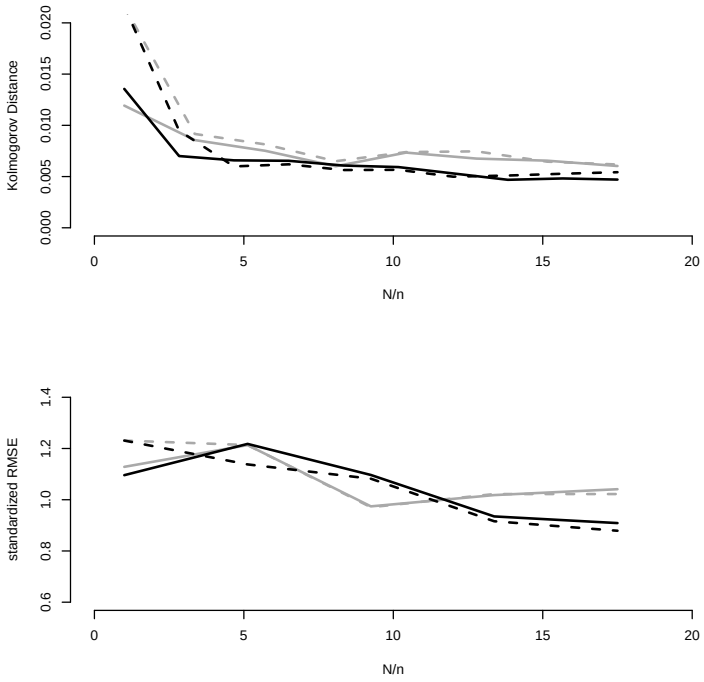

Figure 4: Bias in Matched Samples

7. Bias Reduction through Matching

In this section, we consider how the ratio of the size of the matching pool to the target sample size affecs the amount of bias reduction. If the matching ratio is one, then the matched sample is identical to the unmatched sample and there is no bias reduction. As the matching ratio increases, matching more closely replicates the target distribution of the matching variables, which should tend to reduce bias.

After drawing a stratified target sample of $n = 2000$ respondents from the frame, matched samples were created using nearest neighbor matching with a Mahalanobis distance metric with two sets of variables: a base set of age, race, gender, education, and geographic region; and an augmented set containing the base set of variables as well as income, employment status, and marital status. The pool of respondents from which the matched sample was selected is a simple random sample of the CCES dataset. The size of the pool ranged from $N = 2000$ respondents, so that $N = n$, to $N = 35,000$ respondents. Each matching procedure was replicated five times. The resulting 190 matched samples[8] were weighted using an estimated propensity score model utilizing the same variables as in the matching distance metric.

In Figure 4, the gray lines refer to estimates using the base set of matching and weighting variables, while the black lines refer to estimates using the full set. Dashed lines refer to unweighted estimates, and solid lines refer to weighted estimates. As can be seen from the top panel, which shows the Kolmogorov distance between the sample empirical distribution function of the matching variables, a matching ratio of about five appears sufficient to match the distribution of covariates. With

[8]There was a base and an augmented set of matching variables, each matching scheme was replicated five times per pool size, and 19 different pool sizes, $\{1000, 2000, 3000, \ldots, 10,000, 12,000, \ldots, 20,000, 23,000, \ldots, 35,000\}$, were used.

weighting, a matching ratio of about three is sufficient.

The bottom panel of Figure 4 shows the standardized RMSE for the Obama vote share in states with thirty or more respondents. If all bias is removed, the standardized RMSE should be about one. Generally, the full set of matching and weighting variables is more effective at removing bias (though the base set performs adequately) and a matching ratio of about ten is necessary to achieve nearly complete bias elimination.

8. Conclusions

The effectiveness of matched sampling for removing selection biases from an opt-in panel depends upon the size of the panel, its overlap with the target population, and the ignorability of panel participation conditional upon a chosen set of covariates. The validity of these assumptions will differ depending upon the chosen area of application. In the 2008 U.S. Presidential election, logistic regression estimates of vote choice from different sample sources indicate that all (or at least all but one) of the samples were subject to at least some form of nonignorable selection. However, the magnitude of the differences was generally modest and evidently no greater for the opt-in Internet sample as the more traditional RDD- based methods.

Because of imperfect matching, it is still necessary to weight the sample after matching, even if the ratio of the pool of available observations to the target sample is small, especially if the matching ratio is less than five. With or without weighting, the sampling distribution of the matched sample (with a 10:1 matching ratio) exhibited less bias than the RDD telephone samples and much less bias than the RDD-based Internet panel. In all cases, the sampling distributions for both opt-in and RDD samples were approximately normal. The estimate of standard errors using a variance inflation factor were close to the empirical standard errors.

REFERENCES

Cochran, W.G. (1973), *Sampling Techniques*, 3rd ed., Wiley, New York.

Ireland, C.T., and Kullback, S., (1968), "Contingency Tables with Given Marginals", *Biometrika* 55, 179-188.

Little, R.J.A., and Rubin, D.B., (2002), *Statistical Analysis with Missing Data*, 2nd ed., Wiley-Interscience, Hoboken, NJ.

Malhotra, N., and Krosnick, J.A., (2007), " The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples", *Political Analysis* 15, 286-323.

Manski, C.F., and McFadden, D., (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis" in Manski, C.F., and McFadden, D., eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge.

Rivers, D., (2007), "Sampling for Web Surveys", Paper presented at the *Joint Statistical Meetings*.

Rosenbaum, P.R., and Rubin, D.B., (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 41-55.

Vavreck, L., and Rivers, D., (2008), "The 2006 Cooperative Congressional Election Study", *Journal of Elections, Public Opinion and Parties* 18(4), 355–366.