THE GOOD CITIZEN BIAS:
DOES RANDOM-DIGIT DIALING OVER-INCLUDE UNLIKELY VOTERS IN OPINION SURVEYS?

Kenneth Stuart Loman
B.A., University of California, Santa Barbara, 1988

THESIS

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF PUBLIC POLICY AND ADMINISTRATION

at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

SPRING
2011

THE GOOD CITIZEN BIAS:
DOES RANDOM-DIGIT DIALING OVER-INCLUDE UNLIKELY VOTERS IN OPINION SURVEYS?

A Thesis

by

Kenneth Stuart Loman

Approved by:

_____, Committee Chair
Robert W. Wassmer, Ph.D.

_____, Second Reader
Edward (Ted) L. Lascher, Ph.D.

_____
Date

Student: Kenneth Stuart Loman

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

_____, Department Chair _____
Robert W. Wassmer, Ph.D. Date

Department of Public Policy and Administration

Abstract

of

THE GOOD CITIZEN BIAS:
DOES RANDOM-DIGIT DIALING OVER-INCLUDE UNLIKELY VOTERS IN OPINION SURVEYS?

by

Kenneth Stuart Loman

Voter opinion surveys help frame American debate on matters of public policy and can influence legislators' decisions. Voter surveys generally use one of two sampling methods. Voter file sampling includes information about respondents, but may lack unlisted phone numbers. Random digit dial (RDD) sampling avoids this problem, but lacks information about respondents. Consequently, RDD surveys identifying likely voters based on prior voting behavior must rely on information from respondents. However, the literature notes that over reporting of prior voting behavior is widespread. Thus, RDD likely voter surveys risk including inappropriate respondents.

This thesis explores three major questions using a survey of 800 registered voters in Contra Costa County, California. First, is it feasible to predict over reporting using information generally collected during RDD surveys? Second, are over reporters different demographically than true likely voters? Third, does it matter – do the two groups differ on matters of public policy? I found that large numbers of respondents

over reported voting history. Multiple regression of the survey data provided little support for the feasibility of predicting over reporting. However, statistical analysis showed that over reporters are significantly different from true likely voters, both demographically and in policy preferences. Consequently, RDD surveys are not likely to reflect the attitudes of true likely voters, and consumers of such surveys risk making policy and law with bad information.

_____, Committee Chair
Robert W. Wassmer, Ph.D.

_____
Date

ACKNOWLEDGMENTS

I would like to express my gratitude to Elaine Hoffman of EMH Research and Bob Proctor of Statewide Information Systems, without whose support and collaboration this research would not have been possible.

I would also like to thank Professors Rob Wassmer, Ted Lascher, and Mary Kirlin whose faith and confidence helped me into the program, and whose guidance and support helped me through it.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

Chapter 1

INTRODUCTION

The primary purpose of this thesis is to explore the predictability of inclusion of ineligible respondents in public opinion surveys using random-digit dialing sampling methodologies. Specifically, the analysis explores several regression techniques to evaluate whether or not various demographic factors affect the likelihood that respondents will over-represent their past voting history, leading to potentially erroneous inclusion in the survey sampling frame. In each case, the dependent variable is a dummy variable indicating whether or not, or the degree to which, the respondent overrepresented their voting history for a set of prior elections.

The secondary purpose of this thesis is to explore two corollary questions: 1) Are likely voters who have been inappropriately identified due to their incorrect reporting of prior voting behavior different then likely voters who have been correctly identified from voting records. 2) Does it really matter – do overreporting likely voters actually differ on policy or political preferences?

Answers to these questions have value for several reasons. People, including policy makers, pay attention to polls. Polls are informative about how candidates and issues are faring in the time leading up to an election. Additionally, they inform policy debates by making public officials aware of public opinion on matters of public policy. This begs the question of the accuracy of public opinion polls, and raises the profile of otherwise arcane questions of survey methodology.

The remainder of this first chapter provides some general background and support for the argument that the methodologies used to gather public opinion data are relevant and important for users of polling data in understanding how to assess the accuracy of reported results. I discuss how polls affect public policy and are central to theories of democracy, and the importance of likely voters and getting an accurate picture when researching public opinion. I then present the methodological issue central to this thesis, comparing use of voter file sampling with random digit dialing, and conclude with an overview of the remaining chapters.

Polls Affect Public Policy

Since the infamous "Dewey Defeats Truman" debacle in the 1948 presidential election, public opinion polling has risen to a dominating position in American politics. We are all familiar with pre-election polling telling us who is winning the horserace of weekly tracking polls, how ballot measures would fare "if the election were held today," and how much of the public thinks the state is on the right track. Because California is a state with initiative, referendum, and recall processes, policy makers have an incentive to pay close attention to public opinion on policy issues, especially to those most likely to vote. In the November 2010 election, for example, local governments placed a measure before voters to protect local revenue sources, in direct response to legislative budget decisions adversely affecting local government financing.

Non-partisan, non-profit organizations such as the Pew Research Centers and the Public Policy Institute of California have made public opinion polling a major part of the research they provide to the public and policy makers, in the first case with a

national perspective and in the second case with a California perspective. The Field Poll and the Los Angeles Times poll also provide non-partisan public opinion polling for policy makers and the public. Generally these organizations publish reports through news media, and sometimes hold briefings for the press and policy makers.

One indicator of the high profile role polling plays in policy formation is the level of complaint. As early as 1994, for example, former Representative Ron Klink (D-PA.), after a meeting of a House labor-management subcommittee, lamented the growth of "government by polling":

> Every member that got up to talk about whether a benefit should be in the package or not was quoting some poll. Every member has some half-assed poll of his own district, and members use them whatever way they want. Everyone is using some poll or another in every discussion (Schribman).

Often the link between public opinion polling and policy development or legislative action is not so clear. For example, one could argue that Californian's strong support for the state's climate change law (the Global Warming Solutions Act of 2006), has allowed the state's Air Resources Board, charged with implementing the act, to develop more aggressive regulations and implementation targets than might be the case otherwise. While there may not be documentation of a direct link, public support for such policies has been evident in poll results (Baldassare, *et. al.*, 2010), and the recent defeat at the polls of Proposition 23 on the November 2010 ballot, which would have delayed implementation of the act.

Congressman Klink's laments may live on as well at the national level, and are shared by those on the other side of the aisle – as exemplified by a 2006 story in The Weekly Standard titled: The Coming Immigration Deal; Congress Will Follow the Polls. That story noted that public opinion was shifting in support of so-called "amnesty" for illegal immigrants in the U.S. Ironically, the latest iteration of that debate, the "Dream Act," failed during the lame duck congress following the November 2010 election amid polling suggesting that public opinion had shifted in the other direction (PR Newswire, 2010).

Polls are Central to Theories of Democracy

In a theoretically "pure" democracy, there would be no difference between public opinion and public policy; the electorate would debate and decide any question of policy.

Elected representatives, such as in a democratic republic like most levels of government in the United States, have considerable freedom to decide how to act in the public's interest. More and more, public opinion polling provides a method of informing public officials of the sentiment of the electorate.

Researchers Celinda Lake and Jennifer Sosin explore this issue further in their 1998 National Civic Review article "Public Opinion Polling and the Future of Democracy" (Lake & Sosin, 1998). In their view, the explosion of political polling:

> Starkly reveals two fundamentally differing visions of how representative democracy should work. In one vision, representatives are elected to give direct voice to the people's preferences. In the other, representatives serve more as

delegates than representatives; they are invested with the trust to exercise their own judgment.

On the other hand, attempting to understand the will of the electorate through polling is not without danger. Keeter (2008), Director of Survey Research for the Pew Research Center for the People and the Press, suggests that "at a deeper level, the unease about polling grows out of fears about its impact on democracy". He points to criticism that early projections of Ronald Reagan's victory in 1980 may have discouraged some west coast voters from going to the polls to vote for Jimmy Carter.

Keeter's early projection example highlights concern about how the rise of public opinion polling affects the way news is presented. The Pew Center's President, Andrew Kohut (2009), noted a 2008 observation by former CBS News pollster Kathleen Francovic commenting on the effect of polling on new coverage: "polls have become even more important and necessary to news writing and presentation, to the point where their significance sometimes overwhelms the phenomena they are supposed to be measuring or supplementing."

While the idea that polls as process stories may overshadow substantive coverage of news is a serious issue, an even more fundamental issue underlying Keeter's comment is that polling places public opinion front and center in any significant policy debate. News organizations use polling both because the poll's snapshot of public opinion may be a story in itself, and also because it provides a benchmark for stories assessing the performance of public officials. Campaigns use them to hone messages in the light of public opinion, as the polls reveal it.

The Importance of Likely Voters

Content and ideology aside, these articles show the impact that voter opinion polls have on how policy debates are covered by the media, and how these debates are framed to begin with. Such an influential role, of course, begs the question: are these polls really accurate? The 2008 version of "Dewey Defeats Truman" was the chorus of pollsters predicting the victory of Barack Obama over Hillary Clinton in the New Hampshire Primary. The margin of victory predicted by this chorus averaged eight percentage points (Keeter, 2008). Clinton won with 39 percent to Obama's 36, a swing of 11 percentage points from the chorus' prediction. Emblematic of the issue, the headline of a story published by New Hampshire Public Radio read: *Pollsters Wonder How They Got It Wrong on Hillary Victory*.

Scott Keeter provides a brief historical perspective on the polling errors from the New Hampshire primary:

> The New Hampshire debacle was not the most significant failure in the history of public-opinion polling, but it joined a list of major embarrassments that includes the disastrous Florida exit polling in the 2000 presidential election, which prompted several networks to project an Al Gore victory, and the national polls in the 1948 race, which led to perhaps the most famous headline in U.S. political history: "Dewey Defeats Truman" After intense criticism for previous failures and equally intense efforts by pollsters to improve their techniques, this was not supposed to happen. (Keeter, 2008)

In the classic case of the "Dewey Defeats Truman" prediction, the polling error was simple. In 1948, the distribution of telephones was economically skewed such that more wealthy people, who were more likely to be Republicans, had phones than less wealthy people, who were more likely to be Democrats. The telephone poll conducted by the Chicago Tribune failed to correct for this bias and as a result reported biased, and inaccurate, results. In New Hampshire in 2008, there were several likely problems facing pollsters. One was whether or not their samples were somehow biased toward one candidate or the other. Another problem was determining who was likely to actually vote. Yet another problem was that the campaigns were aggressively fighting over potential voters even as the pollsters were attempting to measure opinions. All of these problems were exacerbated by the fact that the universe in which they were polling was relatively small, creating various technical problems for sampling and analysis.

Getting an Accurate Picture

Starting from the premise that the basic science of polling is sound, my purpose here is to explore one corner of this question, focused on how researchers select respondents for inclusion in public opinion surveys, specifically surveys of likely voters.

Identifying and avoiding ineligible respondents is of critical importance to researchers, both for methodological reasons and because of the cost of conducting surveys. Screening out ineligible respondents from a survey sample can significantly increase the cost of the survey. The perfect sample, then, would include only eligible respondents. In a survey of likely voters, this would mean that each potential respondent in the sample

would be a registered voter who meets the definition of a "likely voter," however the particular researcher defines that group. This leads to the debate underlying the research presented here.

Voter Files versus Random Digits

In my experience managing data collection for voter opinion surveys, I found clients to be divided between sampling of a state's "voter file," or list of registered voters, and random-digit-dial (RDD) sampling, which in essence involves calling randomly constructed telephone numbers. Regardless of the sampling methodology used, clients' based their definitions of likely voters on respondents' past voting histories.

Generally, voter files include a potential respondent's past voting history as well as registration status. This means the sampling frame can be limited to likely voters prior to the selection of a random sample for use in a survey. A key benefit of this is that the researcher knows that each potential respondent is eligible to participate in the survey. This eliminates the cost of screening out ineligible respondents, along with any uncertainty in planning the research project's budget arising from uncertainty about the incidence of eligible respondents in the survey sample.

The debate arises because voter file sampling has a potential Achilles heel – not all registered voters include their phone numbers in their voter registration information. "Phone match" services, which use data mining techniques to find phone numbers from public listings and other sources, can improve the quality of the list, but they cannot make it perfect. The key question is whether voters with unlisted phone numbers behave

differently then voters with listed phone numbers. If the answer is yes, then voter file sampling is inherently biased in a manner similar to the "Dewey Defeats Truman" error.

Random-digit-dial sampling provides a solution to the problem of unlisted phone numbers. Because RDD samples include randomly generated phone numbers, they are likely to include a representative sampling of both listed and unlisted phone numbers. The problem with RDD sampling is that the researcher has no information about a potential respondent until an interviewer speaks with them. Most likely, the incidence of eligible respondents in the sample is the same as the incidence of eligible respondents in the general population, and the researcher must include within the project budget the cost of screening out ineligible respondents. To do so, interviewers ask respondents to answer screening questions to gather their registration status and voting history to determine if they are likely voters, and thus eligible for inclusion in the survey.

Random digit dialing is a more expensive process, since time is spent screening potential respondents, but is valuable if unlisted phone numbers reach likely voters who are different from those with listed numbers. This argument is vulnerable, however, to the problem of inappropriate inclusion of ineligible respondents who overrepresented their voting history in response to screening questions.

The central research question of this thesis is to explore factors affecting the propensity of survey respondents to over-represent their voting history when asked screening questions to determine their inclusion in the survey's sampling frame. Such questions are asked at the beginning of voter surveys using random-digit dialing

methodologies, since this information about a prospective respondent is not known. The research reported in this paper is consistent with that discussed in the literature section, and shows that people do over-represent their voting history, which I call the "Good Citizen" bias. For example, in the data set used for this analysis 29.4% of survey respondents said they had voted in the November 1998 general election when in fact they had not. In contrast, only 6.4% of respondents under-represented their voting history for that election.

As with respondents that have unlisted phone numbers, inclusion of misreporters in the sampling frame is a problem if those respondents are different from respondents who accurately respond to screening questions. The example of asymmetric misreporting mentioned above suggests that they are. The literature discussed below also indicates that likely voters are different both from non-likely voters and from non-voters in significant ways, including their opinions on policy issues. An inaccurate sampling frame that includes overreporters among likely voters could therefore bias a survey's results. As a result, policy makers using such a survey as an indicator of the attitudes of the electorate, or as a predictor of potential voting behavior, might be basing their assessments on inaccurate information.

Baldassare (2006) highlights the political differences between likely voters and nonvoters:

> Likely voters are deeply divided about the role of government, satisfied with initiatives that limit government, relatively positive about the state's elected leaders, and ambivalent and divided along party lines on ballot measures that

> would spend more on the poor. In contrast, the state's nonvoters want a more active government, are less satisfied with initiatives that limit government, are less positive about elected officials, and favor ballot measures that would spend more on programs to help the poor.

It is clear even from the broad strokes of this analysis that surveys assessing the attitudes of those likely to express their political will at the ballot box run the risk of presenting significantly different results depending on how accurately prospective respondents are screened.

Thesis Chapters

Following this introduction, this thesis includes four additional chapters. Chapter 2, Literature Review, includes a review of selected academic literature related to the thesis question, a discussion of position of this research within that literature, and an overview of the analytical model and included variables. Chapter 3, Methodology, describes the survey methods used to collect data for this analysis, and includes a detailed discussion of analytical methods used in this analysis, as well as discussion of possible sources of errors. Chapter 4, Results, includes a discussion of the results of the analysis and range of likely errors from the sources described in chapter 3. Chapter 5, Conclusions and Implications, provides a summary of the research reported in this paper and the findings of the research, discussion of the implications of those findings for the academic literature as well as for practical application, along with a discussion of possible directions for future exploration of the topic.

Chapter 2

LITERATURE REVIEW

Research into electoral behavior and public opinion comprise a broad field of academic inquiry, much of which is based on survey research. Validation of survey data challenges researchers to identify types of errors and tests for those errors, and to develop research methods that avoid such errors in the first place. One critical, and fundamental, source of error is definition of the survey sampling frame. As discussed in the previous chapter, and in further depth below, likely voters are different both demographically and politically from others who might be included in public and voter opinion surveys. As a result, it is important for researchers to take steps to ensure the accuracy of the sampling frame on which assumptions about the results of research are based.

In documenting the widespread nature of vote over-representation by survey respondents (Belli et al, 1999; Freedman and Goldstein, 1996; Presser and Trougott, 1992; Presser, 1990), the academic literature supports the need to improve understanding of such behavior. This chapter explores the literature related to: differences between likely voters and others and the prevalence of vote overreporting in surveys; reasons for overreporting such as social desirability bias (my good citizen bias), the role of memory, and problems with it as an explanatory variable; understanding who overreporters are, explanatory variables associated with overreporting and models overreporters; and finally, the variables included in this analysis.

Likely Voters are Different than Other Groups

The Public Policy Institute of California (PPIC) used data from its PPIC Statewide Survey to compare profiles of likely voters with infrequent voters and those not registered to vote. The PPIC study found that likely voters differ across several key dimensions. California's likely voters are more conservative, geographically skewed (slightly) toward the San Francisco Bay Area over Los Angeles County, "disproportionately white," and "more affluent, more educated, older" than infrequent voters or those not registered to vote (PPIC, 2010).

As discussed in the previous chapter, California's likely voters differ politically as well as demographically. Some specific examples relate to Californians' attitudes and preferences related to environmental and energy policies. For instance, while 59% of both all adult residents and likely voters are opposed to "allowing more offshore drilling off the California coast," the groups have different attitudes towards "building more nuclear power plants at this time": 53% of likely voters favors the idea compared with only 44% of all adults (Baldassare, *et. al.*, 2010).

Nationally, the differences between likely voters and others are similar. The Pew Research Center for the People and the Press compared likely voters with nonvoters. Their profile focused on nonvoters, noting that "turnout in midterm elections is typically less than 40% of the voting age population" and that likely nonvoters "constitute a majority of the American public." Demographically, the Pew profile found that "nonvoters are younger, less educated and more financially stressed than likely

voters." Politically, "nonvoters are significantly less Republican in their party affiliation than are likely voters, and more supportive of an activist federal government."

Prevalence of Over Reporting

Consistent with the fundamental finding of my research, vote over-reporting is ubiquitous (Presser, 1990). Several key studies have quantified vote overreporting behavior by survey respondents. Parry and Crossley (1950) conducted one of the first efforts to quantify and analyze vote overreporting in 1949. They compared survey responses to public records to validate respondents' "registration and voting in six city-wide Denver elections held between 1944 and 1948." (p. 70). They found that 16 percent of respondents overreported registration (versus two percent underreporting) and between 13 percent and 28 percent of respondents inaccurately reported voting in one of the six specific elections (compared with three percent or less underreporting).

Comparing survey responses following a national election with public election records, Presser and Traugott (1992) found that 13 percent of respondents inaccurately recalled having voted (they did not identify underreporters). Exploring more broadly, Belli, Traugott, and Beckman (2001) examined data from the National Election Studies for seven national elections and found vote overreporting to range from 7.8 percent to 14.2 percent, with an average of 10.2, of respondents. This compared with an average of 0.7 percent of respondents who underreported voting.

Good Citizens and Opinion Polls

The issue of accurately including respondents in a survey is fundamental. Basing conclusions on information revealed by respondents, however, is only as accurate as the

information provided. If definition of likely voters is based on prior voting behavior, this poses a serious problem. Understanding why respondents may provide inaccurate information can lead to better screening of inappropriate candidates for participation in a survey.

Connelly and Brown (1994) explored issues related to gathering information at the individual level and determined that "the reasons for misreported data include both memory-recall errors and social desirability bias." They include a useful discussion of the later:

> Social desirability bias (sometimes called prestige bias) refers to the tendency of the respondent to over- or underestimate participation in an activity or strength of an attitude because of the perceived status given a particular answer. … Others have documented its existence where validation was possible in situations such as reported voting behavior, contributions to charitable organizations, crime reports and so forth.

As Presser (1990) points out more generally, "vote overreporting has been found in every major validation study." He attributes this to social desirability bias on the part of respondents: "The problem of vote overreporting is presumably due to the fact that people like to see themselves as good citizens or, more generally, to present themselves in a socially desirable light" (ibid, page 587).

The Role of Memory in Over Reporting

In addition to social desirability bias, Connelly and Brown (1994) explored the role of memory failure as a causal factor in misreported data. Their analysis assumed

that memory recall errors worked both ways. They allocated all underreporting as memory error (since it is not socially desirable), and then subtracted that amount from the total over-reported, and allocated the remainder to social desirability bias.

Belli, et. al. (1999) note that research does not support one theory that a memory of a prior experience of voting could take the place of a failed memory of a recent election and cause a respondent to report voting when in fact he or she had not voted. This is consistent with the findings of Presser and Traugott (1992) that respondents generally attempt to answer truthfully about voting.

Additionally, Belli, et. al. take the analysis of Connelly and Brown a step further, suggesting a synergistic relationship between memory and social desirability:

> Instead of attempting to attack either social desirability or memory failure separately, we consider overreporting to be a result of their combined influences… Hence, whenever respondents do not precisely remember that they did not vote in the last election, the social desirability of voting is seen to bias respondents to overreport.

Interestingly, this is consistent with psychological research documenting the malleability of memory itself, such as the power of advertising to alter subjects' actual memories of product experience – for example the taste of orange juice (Braun, 1999).

A critical implication of a synergistic or malleable theory of memory is that the passage of time is less likely to have value as an explanatory factor specifically in vote overreporting. It is reasonable to assume that the more time that has passed between an election and a respondent's attempt to recall their behavior, the less likely it is that the

respondent will provide accurate information, opening the door to synergistic effects that might blur causation.

Problems with Memory as an Explanatory Variable

The role of memory is somewhat ambiguous, and perhaps not relevant as an explanatory variable (Presser and Traugott, 1992, page 79). In addition to theoretical challenges associated with the role of memory as an explanatory variable, there are several practical problems as well. For example, the data used for this analysis is composed of 800 interviews, during which each respondent was asked to recall voting in each of four prior elections. Including a variable for the length of time between the interview and each prior election would be simple if the level of analysis were the specific recall report for each election. However, disaggregating each respondent's data into four separate records poses significant problems in modeling the variables associated with individual respondents' characteristics. On the other hand, using the respondent as the level of analysis requires aggregating responses to the four election voting behavior questions into a dichotomous or scale variable (depending on the type of analysis), making it extremely difficult to control for elapsed time or other memory proxies.

Another challenge with including memory, or some suitable proxy, in the analyses discussed here is the practicality of defining memory related variables useful for identifying potential respondents in a random digit dial sample who should be excluded from a survey on the grounds that they are likely outside the sampling frame of likely voters; that they are not likely to be likely voters. This again goes to the key

thrust of this paper, that identification of likely voters in a survey for campaign or policy use requires a model that can accurately screen out inappropriate respondents, rather than a model designed to explore the reasons for respondents' behavior. Because I am focusing this research on the practicality of predicting overreporters from the types of data generally gathered in commercial surveys, I am focusing on the respondent as the level of analysis and not including variables related to memory in my regression analysis. However, it should be possible to see if the data support any general conclusions regarding memory by simply comparing the rates of misreporting across the four elections included in the survey.

The theoretical and practical problems with memory discussed above suggest that a better approach in attempting to identify overreporters in survey samples might be to focus on characteristics common to respondents most likely to provide inaccurate information.

Who are the Over Reporters?

Unfortunately, modeling the overall characteristics of over-reporters is rather difficult, and studies are conflicted. Freedman and Goldstein (1996) find that respondents who over-report voting more closely resemble non-voters than voters. On the other hand, Presser and Traugott (1992) found mis-reporters "tend to resemble actual voters". They go on to suggest this is because respondents are untruthful about self reporting on other characteristics as well. The implication is that any modeling of voters v. non-voters based on self-reported data is similarly vulnerable. This is refuted partially by their finding that "misreporters are about as informed as validated voters,

casting at least some doubt on the hypothesis that they reported inaccurately about their education or interest" (Presser and Traugott, 1992, page 83).

Given such uncertainty, a reasonable first step is to determine the practicality of identifying over-reporters, using data generally collected during voter-opinion studies. To this end, the model explored here is based completely on data normally collected in the course of such surveys.

The focus of the Belli, Traugott, and Beckman (2001) analysis was a comparison of overreporters to both validated voters and admitted nonvoters, to see whether overreporters as a group were similar to either of the other groups. They found that the three groups:

> Represent basic populations that differ in their characteristics. Overreporters are situated in between validated voters and admitted nonvoters in their age, level of education, and strength of political attitudes. With the exception of age, overreporters are significantly closer to validated voters than nonvoters in these measures. Overreporters are predominantly non-white, and overreporting occurs more frequently the further the election takes place from election day.

With respect to why overreporting occurs, their results were also consistent with their theoretical argument that "overreporting is due to a combination of motivational and memory factors."

Presser and Traugott (1992) also found that misreporters differed from actual voters. They performed regression analyses on ANES data to test the hypothesis that misreporters generally voted; that misreporting was an irregular event. Their analyses

found that not to be the case. For example, in comparing validation of self-reported voting behavior in the 1972 and 1976 national elections, they found that (as mentioned above) 13 percent of respondents overreported having voted in one or both elections. Of those, 88 percent had not voted in either election, and only three percent had actually voted in both elections.

Explanatory Variables

If social desirability bias is present in survey data, then it is reasonable to suppose that the characteristics of mis-reporting respondents would correspond to the characteristics of those who value the behavior being reported. Presser and Traugott (1992) put forth this theory in exploring possible causes for their finding that education is related to misreporting: "the better educated and more interested may feel more pressure to misreport because their naïve theories about politics tell them that they are the kinds of people who vote (or, alternatively, ought to vote)". Comparing the results of regression analyses attempting to predict voter turnout based on self-reported voting history versus validated voting history information, they found that education correlated with the self-reported information but not the validated information.

Other specific characteristics that have been found to be significant are ethnicity and location of residence. Connelly and Brown (1994) found that white respondents were approximately twice as likely as non-whites to over-report having contributed to a wildlife income tax check-off program in New York State. They also found that over-reporting varied by residence location. Those living in "villages" of less than 25,000 were more likely to over-report than those living in "cities" over 25,000 or rural areas.

Building a Model

Attempting to develop a more specified model, Belli, Traugott, and Beckman (2001) developed a regression model that looked at three categories of variables with the potential to predict respondents' overreporting of voting history. They analyzed "social predictors" including age, education, race, and gender; "political attitudes" including degree of political efficacy, caring about the outcome of the election, interest in the campaign, strength of party identification, and expressed knowledge of political individuals or groups; and "contextual variables" including time since the election, election type, and the year of the election. They found age, education, ethnicity, and strength of political attitudes to be significant variables in distinguishing overreporters from validated voters and admitted nonvoters.

One intriguing aspect of the model developed by Belli and his group exemplifies the differences between research designed for academic analysis and research designed for commercial use, such as voter opinion polls conducted during a campaign. The data used for their analysis came from the American National Election Studies (ANES), an academic program of the University of Michigan and Stanford University. The specific range of values in the data relating to the amount of time since an election is the number of weeks between the election and when the interview took place. In contrast, voter opinion polls conducted during an election campaign or policy debate may seek a sample of likely voters based on potential respondents' participation in elections over a time span of several years. If that sample selection process relies on information self-

reported by potential respondents, then levels of misreporting may be significantly larger compared with those identified in the academic data.

The Place of this Research in the Literature

This research fills a gap in the literature regarding the practical application of theoretical knowledge of voter behavior to existing practice in the voter opinion research industry. While the professionalism of industry practitioners doubtlessly includes efforts to keep abreast of new learning in the field, changes in practice require clear evidence that there is, in fact, a problem and a demonstrably cost-effective solution.

Two critical factors in the cost of conducting voter opinion research, the length of the interview and the incidence of eligible respondents in the sampling frame, are both affected by the complexity of the screening process. Improving screening with data gathered by current question sets, therefore, has a higher likelihood of implementation than methods that might require new and possibly longer screening sets. Conversely, a clear understanding of the risks of not improving screening may change the relative assessment of sampling methodologies by pollsters.

This research addresses both aspects of this gap. My primary focus is on the feasibility of using data gathered in general practice to improve screening. Secondarily, I expand on that feasibility study to assess the risks of including inappropriate respondents in a survey who could damage the quality of its results.

Chapter 3

METHODOLOGY

This chapter includes discussion of the research design and analytical model, the target population and sampling methodology, and the data used for this thesis. The research design and analytical model involves three interrelated research questions regarding the predictability of inappropriate identification of survey respondents as likely voters, and the consequences of inclusion of inappropriate respondents for survey results. I specify appropriate analytical methods for evaluation of the three research questions, including definition of dependent and explanatory variables, and the expected impact of explanatory variables. The section on data includes the definition of, and rationale for, the target population, and a description of the stratified voter file sampling method used. This final section also includes identification of sources and descriptive statistics for all variables used in this analysis.

Research Design and Analytical Model

This section includes an overview of the basic research design for the three research questions and analytical methods for each, including specification of regression models for analysis of research question #1 and bivariate analyses for comparison of overreporting likely voters and non-overreporting likely voters for analysis of research questions 2 and 3.

In exploring the predictability of the behavior of survey respondents in studies using Random Digit Dialing sampling methods, it is necessary to focus on the kinds of data readily available to survey researchers who use such methods. Consequently, this

analysis uses a quantitative approach to analyze data gained via responses to questions normally asked during such surveys. These data are used to explore three interrelated questions. First, is it feasible to use the information generally collected during voter opinion surveys to predict overreporting of voting histories by respondents who might be identified inappropriately as likely voters? Second, are vote overreporters who are identified as likely voters in this study different from likely voter respondents who do not overreport their voting histories, as suggested by the literature? Third, does it matter – do vote overreporters identified as likely voters actually differ on policy preferences?

The basic research design for the first research question compares survey respondents' answers to standard RDD screening questions with their actual voting records. To facilitate this validation of survey responses, the survey used a voter file sample that included voting history information derived from official records. The basic research design for both the second and third research questions compares two populations (overreporting likely voters and non-overreporting likely voters) across two sets of comparison variables. The background comparison data for this analysis was either included in the sample or gathered from respondents during the interview. The following sections specify the analytical methods for exploring each of the three research questions.

Research Question #1: Feasibility of Predicting Vote Over Reporters

The primary question driving this thesis is the feasibility of using information generally collected during voter opinion surveys with random digit dialing methodologies to predict overreporting of voting histories by respondents who might be

identified inappropriately as likely voters. I use variations of a basic regression model to explore this question from slightly different perspectives. These variations include a logistic regression version exploring a dichotomous aspect of the research question: *Did* the respondent overreport, as well as exploration of the functional forms of a linear regression version exploring a scalar aspect of the research question: *How much* did the respondent overreport.

Variations on a Dependent Variable

The dependent variables for each version of the regression analysis for this thesis derive from validation of responses to four questions asking respondents whether they had voted in four prior elections. Those four elections are the November 1992 American Presidential election, the November 1994 California Gubernatorial election, the November 1996 Presidential election, and the November 1998 Gubernatorial election. The dichotomous dependent variable indicates whether the respondent overreported for any of the four test elections: respondents are coded with a 1 if they over reported, and a 0 if they did not. I use logistic regression to explore the effects of explanatory variables on this variation of the dependent variable. The scalar variation of the dependent variable indicates the number of elections (from 0 to 4) for which the respondent overreported. I analyze this variation by exploring the effects of the explanatory variables using different functional forms of linear regression.

Including both questions in this analysis provides different perspectives on the data – explanatory variables may show their effect more in one analysis than in the other. For example, linear regression may be more sensitive to an explanatory variable

with a graduated effect than logistic techniques. Alternatively, logistic regression may identify an explanatory variable with a dichotomous effect, which might not stand out in a linear regression analysis.

Causal Categories and Proxy Variables

In general, and building on the literature, I propose that a respondent's propensity to over-represent his or her voting history is a factor of four broad causes: the respondent's socio-economic status, political outlook, memory, and personal demographics. Specifically, Propensity (to over-represent) = f (Socio-economic Status, Politics, Memory, Personal Demographics). In selecting specific variables for these categories, I have limited this analysis to those generally used by current voter studies, and am omitting proxies for memory (as discussed in Chapter 2). The specific variables used in each broad causal category are as follows.

Variables related to respondents' Socio-economic Status are employment status, level of education, whether they own or rent their home, the number of years the respondent had lived in the target community (Contra Costa County), and their household income. Collecting data regarding household income is problematic, both because respondents may not know the exact amount, and because they may be reluctant to share such information. As a result, this variable is often structured categorically, as it is here. This allow respondents to select a category within which they believe their household income lies, without asking them to share more specific, and private, information.

Because employment status, level of education, and household income are categorical in nature, with unequal category definitions, I convert them into sets of dummy variables to indicate respondents' inclusion in a particular category. To avoid perfect collinearity, the modal category for each variable group is omitted from the analysis. The omitted category for employment is "full-time," for education level, "college graduate," for household income, "Over $100,000," for age, "40 – 49," and for ethnicity, "White/Caucasian."

Variables for Politics are the respondent's party registration and the respondent's political philosophy. Political philosophy is measured by asking respondents to describe themselves as where they fall on a five-point scale, where one end of the scale represents Very Conservative and the other end of the scale represents Very Liberal. Variables related to Personal Demographics are the respondent's age, gender, and ethnicity. Because the rationale for using random digit dialing over voter file sampling relies in large part on the concern that voters with unlisted phone number are different from voters with listed phone numbers, I include among the variables for Personal Demographics a dummy variable indicating whether the respondent has an unlisted phone number.

Specification of Regression Models

Exploring both dichotomous and scalar variations of the dependent variable requires the use of both logistic regression for the dichotomous version and linear regression for the scalar version.

Logistic regression explores dichotomous version of dependent variable.

For a dichotomous dependent variable such as "did the respondent over report voting history," Ordinary Least Squares (OLS) based regression techniques present problems. First, a linear regression will predict values for the dependent variable outside the possible range of a dichotomous variable. Second, the regression line will also show inaccuracies within the possible range by predicting values other than 0 or 1, the only possible values in a dichotomous variable. Logistic regression corrects these problems by establishing a set of predicted values that follow an "S" curve that remains within the range of possible values and attempts to switch between the poles of the dichotomous dependent variable (0 and 1) as sharply as possible, given the predictive power of the model. The following chapter will explore issues relating to interpretation of results as well as provide results of OLS and Logit methods for comparison.

The logistic regression model to be estimated then, observed across a sample of "N" respondents (where i = 1, 2, 3, …N) is:

(1) Overrep (dichotomous propensity to over report)$_i$ = f(Employment: Part Time$_i$, Employment: Student$_i$, Employment: Homemaker$_i$, Employment: Retired$_i$, Employment: Unemployed$_i$, Education: Grades 1-8$_i$, Education: Grades 9-12$_i$, Education: HSGrad$_i$, Education: Some College$_i$, Education: Post Grad$_i$, Home Owner$_i$, Years in CC$_i$, Home Owner$_i$, Income: 10,000 or Less$_i$, Income: 10,001 – 20,000$_i$, Income: 20,001 – 30,000$_i$, Income: 30,001 – 40,000$_i$, Income: 40,001 – 50,000$_i$, Income: 50,001 – 60,000$_i$, Income: 60,001 – 70,000$_i$, Income: 70,001 – 80,000$_i$, Income: 80,001 – 100,000$_i$, Party: Dem$_i$, Party: Rep$_i$, Party: DTS$_i$, Party: Other$_i$, Political

Philosophy$_i$, Age: 18 – 29$_i$, Age: 30 – 39$_i$, Age: 50 – 64$_i$, Age: 65 and Over$_i$, Gender: Female$_i$, Ethnicity: Black/African American$_i$, Ethnicity: Hispanic/Latino$_i$, Ethnicity: Asian$_i$, Ethnicity: Other$_i$, Phone Unlisted$_i$.

Linear regression explores scalar version of dependent variable.

The scalar version of the dependent variable measures the amount a respondent over reports prior voting history, from zero prior elections to all four prior elections tested. This five-point scale lends itself easily to analysis using OLS linear regression techniques. As mentioned above, this technique offers more sensitivity to explanatory variables with graduated effects. To gain the maximum advantage from this sensitivity, I will explore the effects of different functional forms on analysis of the data, including linear-linear, log-linear, and log-log forms, where appropriate for the type of variable.

To facilitate exploration of different functional forms, three variables (Income, Education, and Age) are treated with an alternative coding scheme from that described above. Instead of being recoded into sets of dummy variables, these variables are left in their original form and treated as a scalar variable where the values in the scale represent the relationships between the categories. Exploring the effect of different functional forms in this manner involves trade-offs between sensitivity to certain explanatory effects, and sensitivity to variance or errors in specification of the data. For example, the categories for Income represent ten-thousand dollar increments with the exception of the final two categories, which represent a twenty-thousand dollar increment and an open-ended category (above $100,000). Similarly, the categorical dummies for Education and Age represent progressive, though not necessarily equal,

intervals. Such a conversion also fails to account for qualitative differences between levels (e.g. is a college degree simply a matter of more years of education?). Even with such specification issues, it is useful to use these analytical tools to explore these data, if only to identify areas of interest for future research.

The linear regression model to be estimated then, observed across a sample of "N" respondents (where i = 1, 2, 3, …N) is:

(2) OverrepAmount (scalar propensity to over report)$_i$ = f(Employment: Part Time$_i$, Employment: Student$_i$, Employment: Homemaker$_i$, Employment: Retired$_i$, Employment: Unemployed$_i$, Education$_i$, Home Owner$_i$, Years in CC$_i$, Home Owner$_i$, Income$_i$, Party: Dem $_i$, Party: Rep $_i$, Party: DTS $_i$, Party: Other $_i$, Political Philosophy $_i$, Age$_i$, Gender: Female$_i$, Ethnicity: Black/African American$_i$, Ethnicity: Hispanic/Latino$_i$, Ethnicity: Asian$_i$, Ethnicity: Other$_i$, Phone Unlisted $_i$.

Specification of Explanatory Variables

In addition to these standard socio-economic variables, I have also included a dummy variable to indicate home ownership and a continuous variable capturing the length of time the respondent has lived in their community. These variables are included for theoretical reasons. Because the "good citizen bias" being explored here is closely related to social desirability bias discussed in the literature, it seems reasonable that indicators of the stability of the respondent's membership in their community might have an impact on how they represent their involvement in that community through voting.

For politics, the variables are party registration and political philosophy. Party registration can be determined through survey responses (in RDD surveys) or through actual registration files (in VF surveys). This study uses the actual voter registration information, recoded into dummy variables. The modal category "Democrat" is omitted from the analysis to avoid multicolinearity. Political philosophy is included to provide a second dimension for analysis of respondents' political views, and is coded in a Likert scale where 1=very conservative, 2=somewhat conservative, 3=moderate, 4=somewhat liberal and 5=very liberal.

Personal demographics are generally captured by asking respondents' age or birth year and ethnicity. Both of these variables are recoded into sets of dummy variables. Omitted modal categories are age=40-49 and ethnicity=white/Caucasian. Gender is usually captured by interviewer observation in RDD surveys (respondents' are generally offended if this question is asked directly) or from the sample in VF surveys. This study uses the actual voter registration information, recoded into a dummy variable coded 1 if the respondent is female and 0 if male.

Finally, a dummy variable is also included to indicate whether the respondent has an unlisted phone number. This is included for two reasons. First, the argument against use of voter file sampling is based on the assumption that respondents with unlisted phone numbers are different than those without. Its inclusion allows comparison of results across this variable (this aspect is not addressed within the scope of this thesis). Second, because it also seems reasonable that this variable may reflect an underlying concern about privacy and sharing information that could impact

respondents' accuracy and truthfulness. For this reason it is included in the model as a control variable.

Expected Impact of Variables

Consistent with the literature on social desirability bias, I expect variables related to socioeconomic status to have a positive impact on a respondent's propensity to over-represent their voting history. Generally, this suggests that the higher one's socioeconomic status, the more one desires to be seen to exemplify desirable characteristics. An alternative theory suggests that the impetus is more about identity, so that higher socioeconomic status inculcates a belief that one is the type of person who exemplifies desirable characteristics, regardless of whether one is seen to do so or not. Finally, the memory synergy theory suggests that such a self perception takes over when a respondent's memory fails to recall a behavior with desirable characteristics, such as voting. Specifically then, employed respondents would have higher propensity, more education would increase propensity, homeowners would have higher propensity than renters, length of residency would increase propensity and propensity would increase with household income.

I am uncertain what the effect of politics would have and include these variables as much as control factors as to explore causation. Regarding personal demographics, I expect age to have an impact such that the older the respondent, the higher their propensity to over-represent their voting history. I expect age to have a lesser effect than the socioeconomic variables, to the extent that age correlates to those variables, or to the extent that it correlates to reduced memory capacity for some older respondents.

The other demographic variables are primary included as control factors, and I expect them to have no effect.

Research Question #2: Are Vote Over Reporters Different?

This section includes an overview of the second research question and appropriate analytical methods, including specification of a new dependent variable and bivariate analytical techniques for comparison of over reporting likely voters with non-over reporting likely voters, including specification of comparison variables, a hypothesis testing model, identification of appropriate statistical analysis methods. Comparison variables in this analysis comprise demographic and psychographic data descriptive of respondent populations.

The second major research question is if vote over reporters in this study are different from respondents who do not over report their voting histories, as suggested by the literature. Because this thesis focuses on the inappropriate identification of likely voters, I limit this second research question to respondents identifiable as likely voters – as might be included, for example, in a proprietary survey used for purposes of influencing legislators.

A dummy variable divides these likely voters into two groups for comparison. One group consists of respondents identifiable as likely voters based on their actual voting history, while the other group consists of respondents inappropriately identifiable as likely voters because they overrepresented their voting history. This analysis is quasi-experimental in nature because, rather than random assignment of respondents to one of the two groups, respondents are assigned based on their behavior.

To explore this second research question, I compare populations of over reporting likely voters and non-over reporting likely voters against the explanatory and control variables described above. This analysis consists of a series of bivariate pairings, each comparing the dummy variable for appropriate and inappropriate likely voters (e.g. did the respondent over report?) against one of the explanatory variables. The series of bivariate analyses as whole identifies any variables on which the two populations differ.

Methodology for Bivariate Analyses

A new variation of the dichotomous dependent variable.

As previously discussed, the dependent variables used in this thesis derive from a series of four validation tests of respondents' over reporting prior voting behavior, that is reporting that they had voted in a test election when in fact they had not. Each respondent can thus be given an "over reporting score," which will fall between zero and four (inclusive). Similarly, each respondent can be given a "voting score" representing the number of the four test elections in which the respondent actually voted.

For purposes of this thesis (and to keep matters simple), likely voters are defined as those respondents who voted in at least 3 of the four test elections. Additionally, respondents who registered to vote after the third test election (November 1996) and voted in the fourth test election (November 1998) are also identified as likely voters. This two step algorithm, applied to respondents' voting scores, yields correctly identified likely voters (based on actual voting history); applied to respondents'

overreporting scores yields inappropriately identified likely voters (based on incorrectly reported voting history).

Explanatory variables.

The explanatory variables used for these bivariate tests are the same as those used for the regression analyses described above, specifically: employment status, level of education, household income, homeowner, length of residence (years in Contra Costa County), party, political philosophy, age, gender, ethnicity, phone unlisted. With the exception of length of residence and political philosophy, each of these variables is categorical, with nominal value categories. Several of these variables (level of education, household income, and age) have a scalar aspect to their categories, but they are treated as nominal rather than ordinal because the categories are not evenly spaced and because I include a category for respondents who refused to answer a particular question. For these bivariate comparisons, I use the original multi-category coded variables, rather than the recoded dummy category variables used for the regression analysis. Length of residence (years in Contra Costa County) is an interval variable, and political philosophy is ordinal.

Formulating hypotheses for testing.

Each comparison tests for difference between the two populations, overreporting likely voters and non-overreporting likely voters, in terms of the explanatory variable in question. More specifically, the tests determine which of two hypotheses is true regarding any difference between the two populations: the "null" hypothesis (H_O) that there is no difference, or the "alternative" hypothesis (H_A) that there is in fact a

difference. These two mutually exclusive hypotheses can be expressed more formally as:

H_O: Overreporting likely voters do not differ from non-overreporting likely voters.

H_A: Overreporting likely voters differ from non-overreporting likely voters.

To determine the validity of the null hypothesis, each pairing of the dependent variable and an explanatory variable is tested to see if the two variables are independent, that is they do not affect each other. If the pairing of variables is independent, then the distributions of overreporting likely voters and non-overreporting likely voters across the tested variable will be the same, or at least within the normal range of error in survey sampling. In this case, the null hypothesis can be accepted that overreporting likely voters do not differ from non-overreporting likely voters on the tested variable. If, on the other hand, the test shows that the two variables are not independent, then the distributions will differ beyond the identifiable sampling error in the data. In this case, the null hypothesis must be rejected, and the alternative accepted that there is in fact a difference between overreporting likely voters and non-overreporting likely voters on the tested variable.

The logic behind the test for independence between the two variables suggests a more operational formulation of the null and alternative hypotheses:

H_O: The dependent variable and tested explanatory variable are independent.

H_A: The dependent variable and tested explanatory variable are not independent.

Identification of test statistic.

For analysis of the categorical (nominal) variables described above, the appropriate test statistic chi square (χ^2) test (Manheim, Rich, & Willnat). Calculation of the chi square statistic begins with a contingency table, or cross-tabulation, for two variables showing the distribution of survey responses for each combination of values for the two variables being tested. The chi square test compares the observed frequencies in the cross-tabulation with the frequencies expected if the two variables were independent. The equation for calculating the chi square statistic is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where: f_o = the frequency *observed* in each cell of the cross-tabulation

f_e = the frequency *expected* for each cell of the cross-tabulation.

Interpretation of the chi square statistic is constrained by the degrees of freedom in the cross-tabulation. The degrees of freedom (*df*) reflect the number of cells in a cross-tabulation (contingency table) whose content are not determined by the previously filled cells (Manheim, Rich, & Willnat). Effectively, the degrees of freedom are equal to one less than the total number of cells in a contingency table, or

$$df = (r-1)(c-1)$$

where: r = the number of categories of the row variable

c = the number of categories of the column variable.

Statistical tables provide test values at different levels of significance (e.g. .001, .01, and .05) for various degrees of freedom.

Two special cases: length of residence and political philosophy.

Unlike the categorical variables in the chi square analyses discussed in the preceding section, the variable for length of residence (Years in Contra Costa) represents values structured in intervals of one year. The variable for political philosophy is ordinal, with data structured as a Likert scale where 1=very conservative, 2=somewhat conservative, 3=moderate, 4=somewhat liberal and 5=very liberal. Consequently, a similar bivariate analysis requires a different statistical techniques for these two variables. Rather than testing length of residence (Years in Contra Costa) against the dependent variable for independence, this analysis compares the distribution of values for the variable Years in Contra Costa for the two populations of overreporting likely voters and non-overreporting likely voters. To do this, I compare the means of the two distributions to determine if they are the same, or if any difference between them is statistically significant (rather than attributable to chance). Similarly, the variable for political philosophy lends itself to a comparison of means for the two populations, as its Likert scale data can be averaged to create populations means for comparison.

The hypotheses for comparing the means of two populations ask whether or not the two means are the same, or more formally

H_O: The two means are equal, or $\bar{x}_1 = \bar{x}_2$

H_A: The two means are not equal, or $\bar{x}_1 \neq \bar{x}_2$

where $\bar{x}$ is the mean of a given population.

The appropriate statistical technique for comparison of two populations across a continuous or interval variable is a comparison of the means of the distribution of

responses for the two populations (Freund & Simon, p. 341). When large samples (greater than 30) for both populations are tested, the test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where: $\bar{x}_1$ = the mean length of residence for overreporters

s_1 = the standard deviation for overreporters

n_1 = the sample size for overreporters

$\bar{x}_2$ = the mean length of residence for non-overreporters

s_2 = the standard deviation for non-overreporters

n_2 = the sample size for non-overreporters.

Assuming that the variable for length of residence in a large sample has a normal distribution, standard statistical tables for normal-curve areas identify the probability that a respondent's length of residence will fall between zero and *z*. Subtracting that probability from .5 yields a one-tailed probability that the difference of means is statistically significant; doubling it provides a *p*-value for the two-tailed probability that one mean is significantly higher or lower than the other mean. If the *p*-value is less than a selected level of significance (α), such as α = 0.05, then the null hypothesis must be rejected. Because the *p*-value essentially represents a confidence interval that two populations means are the same, I do not identify specific level of significance for the test, but rather report the results of the test in comparison to different levels of significance (e.g. .001, .01, and .05) generally used in statistical analysis.

Research Question #3: Does it Matter – Do Vote Overreporting Likely Voters Differ on Policy Preferences?

This section includes an overview of the third research question and appropriate analytical methods. This analysis uses the same dependent variable specified for research question #2 and bivariate analytical techniques for comparison of overreporting likely voters with non-overreporting likely voters (including specification of comparison variables, a hypothesis testing model, identification of appropriate statistical analysis methods). Comparison variables in this analysis represent respondent preferences on policy or political issues.

The third research question is whether overreporting likely voters actually differ on policy preferences – does it really matter if the wrong respondents are included in a sampling of likely voters? The primary research question of this thesis concerns the feasibility of predicting and identifying vote overreporters in a random digit dial survey before they are inappropriately identified as likely voters based on their incorrect reporting of prior voting habits. This prediction problem is predicated on the potential risks of identifying the wrong respondents as likely voters; in particular, the risk that they are not representative of voters in the intended sampling frame, and the risk that they have different policy preferences and bias or skew the results of the survey.

To explore the question of policy differences, respondents were asked questions related to four policy or political issues. The exact wording of the questions and response options is included in Appendix ___. As with the second research question discussed in the preceding section, the two populations of overreporting likely voters

and non-overreporting likely voters are compared in a series of bivariate tests. All of the variables in this section are categorical with nominal value categories and are analyzed using the chi square statistical test described above.

Test variables related to policy preferences.

Issue #1: right direction or wrong track?

The first policy or political issue includes three versions of a standard question asked by many researchers; asking if the respondent felt that things in California "are going in the right direction, or are they seriously off on the wrong track?" The base question was asked three times of all respondents; for "your community," "California," and "The country."

Issue #2: should California dump its new open primary?

The second policy issue included two questions related to California's establishment of an open primary system. At the time the survey was conducted in 1998, California had recently changed its election laws to create an open primary system. Opponents of the open primary were challenging the legality of the new system (it was later invalidated by the courts), and considering ballot measure to repeal or change it.

For the first question, respondents were read pro and con arguments about the open primary and then asked if they supported or opposed the new law (a follow up clarified their responses by asking if they supported or opposed "strongly, or only somewhat?" providing a four point scale for coding responses). The pro argument was that the open primary would "increase voter turnout because it allows voters to vote for

whoever they really want, without being limited by party registration." The con argument was that the open primary "violates the first amendment freedom of association by forcing voters of one party to allow outsiders to influence their choice of their party's candidate, even if those outsiders are from an opposing party." To prevent primacy or question order biases, the pro and con arguments were rotated between respondents, so that half heard the pro argument first, and the other half heard the con argument first.

The second question regarding the open primary included an additional negative argument that neither the Republican nor the Democratic party recognized the results of open primaries – so that "the California primary will be only a "beauty contest" and will not even be counted in selecting party candidates for President." After hearing this additional argument, respondents were asked if they favored or opposed (and strongly or only somewhat) changing the law so that the California primary will "be counted when the national Democratic and Republican parties choose their candidates for President?" The purpose of this negatively biased second question was to see if there was any difference in the changeability of preferences between overreporting likely voters and non-overreporting likely voters.

Issue #3: presidential horse race.

The third policy or political issue includes only a single question asking respondents' preferences in a head to head test of two (at the time) likely nominees for President in the November 2000 election. The two potential nominees were "Texas Governor George W. Bush, Jr., the Republican" and "Vice-President Al Gore, the

Democrat." Interviewers rotated these options to prevent primacy or ordering bias, so that half the respondents heard the potential Republican nominee first, and the other half of the respondents heard the potential Democratic nominee first. Respondents expressing doubt about their preference were coded as "leaning" toward one or the other potential nominee, though this option was not read to respondents.

Issue #4: local school bond measures.

The fourth policy or political issue included a set of four sequential questions asking respondents to indicate their support of or opposition to a local school bond measure, if it meant that their property taxes would be increased by a certain amount. The first question asked if they would support or oppose the school bond measure if their property taxes would increase by $53 per year. After responding, they were asked the same question, if their property taxes were to increase by $42 per year, and so on for the other two amounts of $36 per year, and $27 per year. All respondents were asked all four questions in descending order.

Before respondents were asked for their preferences, they were provided with some information related to the potential school bond measure. For example, they were told that funding from a 9.2 billion dollar state school bond passed the prior year could be used for "facilities construction, to relieve overcrowding and accommodate growth in student enrollment, and to repair older schools, and for wiring and cabling for education technology." Respondents were also told that "some of this funding is available only as matching funds. This means that only school districts which pass qualifying local bond measures will have access to this funding." After hearing this

information, respondents were then asked for their support of or opposition to the local school bond, for each of four decreasing levels of potential property tax increase.

Data

Population and Sampling Methodology

The target population and sampling frame for this research was all registered voters in Contra Costa County, California. Contra Costa County was chosen in part because the demographics of its population were relatively close to those of the entire state. Additionally, focusing on a single county eliminated the need to control for any regional variation that might result from a sample of all registered voters in California.

The sampling methodology used was a stratified interval sample. In interval sampling, an interval n is determined by dividing the total number of registered voters by the number of desired interviews. The list is then sampled by interviewing every n^{th} name. If an interview cannot be conducted with the identified voter, the next name on the list is called. If the list is not treated prior to sampling, then the researcher must establish quotas for key variables, such as party, gender, and age, to ensure that the data collected is representative of the target population.

In a stratified sample, the list of registered voters is sorted, or "stratified", based on the key variables determining "representativeness." Stratification of the list prior to sampling eliminates the need for quotas and produces data that is automatically representative of the target population. This technique works where the target population is large enough so that if a sampled name (falling on the interval) cannot be

interviewed, the next name, or several names on the list are likely to have the same characteristics with respect to the key variables

Data

This thesis analyzes primary data from a survey of 800 registered voters I conducted in December of 1998 in collaboration with Elaine Hoffman of EMH Research and Bob Proctor from Statewide Information Systems. EMH Research, a professional marketing and opinion research company, conducted the interviews by telephone. At the time the survey was conducted, I was employed by EMH Research as a data collection supervisor. In this capacity, I supervised data collection for this survey, as well as collaborated in the development of the survey instrument. The text of the interview questions is included in Appendix __. The voter file sample was provided by Statewide Information Systems, which maintains voter file data and provides samples for professional research firms.

Sources and Descriptive Statistics

The following tables provide details regarding the specific variables used. Table 1 includes descriptions and sources (omitted modal categories are included in this table for comparison). Table 2 contains descriptive statistics of the variables. For those variables that are dichotomous category dummies, the mean provides an indicator of the percentage of respondents in each category. Finally, a matrix of bivariate correlation coefficients for all included variables is included in Appendix __ (due to size). Most of the data consists of respondents' answers to survey questions asking them to select a category that most closely fits their circumstances or opinions. In most cases, including

the comparison or background variables, no definitions of categories were provided to respondents.

Application of the likely voter algorithms described for research question #2 identified 375 overreporting likely voters and 324 non-overreporting likely voters.

TABLE 1: VARIABLE LABELS, DESCRIPTIONS AND DATA SOURCES

Variable Label	*Description*	*Source*
Dependent		
Overrep	Dummy variable to capture whether respondent overrepresented voting history.	Calculated from comparison of survey responses (Questions 1, 2, 3, & 5) and voter file data
OverrepAmount	Scalar variable to capture amount respondent overrepresented voting history (Range 0 – 4)	Calculated from comparison of survey responses (Questions 1, 2, 3, & 5) and voter file data
Independent: Socio-Economic Status		
Employment	Categorical variable indicating respondent's self-identified employment status	Survey data (Question 12)
Employment: Full Time	Dummy variable representing respondent's self-selected employment status	Recoded from survey response (Question 12)
Employment: Part Time	Dummy variable indicating employment status category selected by respondent	Recoded from survey response (Question 12)
Employment: Student	Dummy variable indicating employment status category selected by respondent	Recoded from survey response (Question 12)
Employment: Homemaker	Dummy variable indicating employment status category selected by respondent	Recoded from survey response (Question 12)
Employment: Retired	Dummy variable indicating employment status category selected by respondent	Recoded from survey response (Question 12)
Employment: Unemployed	Dummy variable indicating employment status category selected by respondent	Recoded from survey response (Question 12)
Education	Categorical variable indicating respondent's self-identified level of education	Survey data (Question 13)
Education: Grades 1-8	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)
Education: Grades 9-12	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)
Education: HS Grad	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)
Education: Some College	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)
Education: College Grad	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)
Education: Post Grad	Dummy variable indicating education category selected by respondent	Recoded from survey response (Question 13)

TABLE 1: VARIABLE LABELS, DESCRIPTIONS AND DATA SOURCES

Variable Label	Description	Source
Home Owner	Dummy variable to indicate home ownership as reported by respondent	Recoded from survey response (Question 15)
Years in CC	Interval variable capturing length of residence in Contra Costa county (interval = 1 year)	Captured from survey response (Question 14)
Income	Categorical variable indicating respondent's self-identified 1997 household income category	Survey data (Question 18)
Income: 10,000 or Less	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 10,001 – 20,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 20,001 – 30,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 30,001 – 40,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 40,001 – 50,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 50,001 – 60,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 60,001 – 70,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 70,001 – 80,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: 80,001 – 100,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Income: Over 100,000	Dummy variable indicating respondent's self-identified 1997 household income category	Recoded from survey response (Question 18)
Independent: Politics		
Party	Categorical variable indicating respondent's party registration	Coded from Contra Costa county voter registration files
Party: Dem	Dummy variable to indicate party registration	Recoded from variable for Party
Party: Rep	Dummy variable to indicate party registration	Recoded from variable for Party
Party: DTS	Dummy variable to indicate party registration	Recoded from variable for Party

TABLE 1: VARIABLE LABELS, DESCRIPTIONS AND DATA SOURCES

Variable Label	Description	Source
Party: Other	Dummy variable to indicate party registration	Recoded from variable for Party
Political Philosophy	Categorical variable indicating respondent's self-identified political philosophy	Survey data (Question 16)
Independent: Personal Demographics		
Age	Categorical variable indicating respondent's self-identified age category	Survey data (Question 17)
Age: 18 – 29	Dummy variable indicating respondent's self-identified age category	Recoded from survey response (Question 17)
Age: 30 – 39	Dummy variable indicating respondent's self-identified age category	Recoded from survey response (Question 17)
Age: 40 – 49	Dummy variable indicating respondent's self-identified age category	Recoded from survey response (Question 17)
Age: 50 – 64	Dummy variable indicating respondent's self-identified age category	Recoded from survey response (Question 17)
Age: 65 and Over	Dummy variable indicating respondent's self-identified age category	Recoded from survey response (Question 17)
Gender: Female	Dummy variable to indicate gender	Recoded from Contra Costa county voter registration files
Ethnicity	Categorical variable indicating respondent's self-identified ethnicity category	Survey data (question 11)
Ethnicity: White/Caucasian	Dummy variable indicating respondent's self-identified ethnicity category	Recoded from survey response (Question 11)
Ethnicity: Black/African American	Dummy variable indicating respondent's self-identified ethnicity category	Recoded from survey response (Question 11)
Ethnicity: Hispanic/Latino	Dummy variable indicating respondent's self-identified ethnicity category	Recoded from survey response (Question 11)
Ethnicity: Asian	Dummy variable indicating respondent's self-identified ethnicity category	Recoded from survey response (Question 11)
Ethnicity: Other	Dummy variable indicating respondent's self-identified ethnicity category	Recoded from survey response (Question 11)
Phone Unlisted	Dummy variable to indicate respondent's self-identified unlisted phone number	Recoded from survey response (Question 19)

TABLE 2: DESCRIPTIVE STATISTICS

Variable Label	Mean	Standard Deviation	Minimum	Maximum
Dependent				
Overrep	.70	.458	0	1
OverrepAmount	1.60	1.36	0	4
Independent: Socio-Economic Status				
Employment: Full Time	.5238	.49975	0	1

TABLE 2: DESCRIPTIVE STATISTICS

Variable Label	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Employment: Part Time	.0863	.28091	0	1
Employment: Student	.05	.21808	0	1
Employment: Homemaker	.0713	.25740	0	1
Employment: Retired	.2388	.42659	0	1
Employment: Unemployed	.0288	.16721	0	1
Education Category (Scale)	4.45	1.11	1	6
Education: Grades 1-8	.0113	.10553	0	1
Education: Grades 9-12	.025	.15622	0	1
Education: HS Grad	.165	.37141	0	1
Education: Some College	.2863	.45229	0	1
Education: College Grad	.325	.46867	0	1
Education: Post Grad	.1838	.38752	0	1
Home Owner	.7625	.42582	0	1
Years in CC	22.15	16.062	1	85
Income Category (Scale)	6.4	2.73	1	10
Income: 10,000 or Less	.035	.18389	0	1
Income: 10,001 – 20,000	.0363	.18703	0	1
Income: 20,001 – 30,000	.0688	.25319	0	1
Income: 30,001 – 40,000	.095	.2934	0	1
Income: 40,001 – 50,000	.0738	.26153	0	1
Income: 50,001 – 60,000	.0975	.29682	0	1
Income: 60,001 – 70,000	.0913	.28814	0	1
Income: 70,001 – 80,000	.07	.25531	0	1
Income: 80,001 – 100,000	.0913	.28814	0	1
Income: Over 100,000	.1513	.35852	0	1
Income: DK	.0375	.1901	0	1
Income: Refused	.1525	.35973	0	1
Independent: Politics				
Party: Dem	.4938	.50027	0	1
Party: Rep	.3388	.47358	0	1
Party: DTS	.0938	.29166	0	1
Party: Other	.0738	.26153	0	1
Political Philosophy	2.8922	1.07462	1	5
Independent: Personal Demographics				
Age: 18 – 29	.1288	.33513	0	1
Age: 30 – 39	.2	.40025	0	1
Age: 40 – 49	.2463	.4311	0	1
Age: 50 – 64	.2188	.41366	0	1

TABLE 2: DESCRIPTIVE STATISTICS

Variable Label	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
Age: 65 and Over	.1838	.38752	0	1
Age: NA/Refused	.0225	.1484	0	1
Gender: Female	.5338	.49917	0	1
Ethnicity: White/Caucasian	.7363	.44094	0	1
Ethnicity: Black/African American	.07	.25531	0	1
Ethnicity: Hispanic/Latino	.0663	.24887	0	1
Ethnicity: Asian	.0413	.19899	0	1
Ethnicity: Other	.035	.18389	0	1
Ethnicity: DK/NA/Refused	.0513	.22065	0	1
Phone Unlisted	.2725	.44552	0	1

Chapter 4

RESULTS

This chapter includes results of my analysis of survey data described in Chapter 3. The following sections include a description of the amount of over reporting of prior voting behavior by respondents, and descriptions and results of the specific analyses exploring each of the three research questions of this thesis. The primary research question is the feasibility of predicting vote over reporters in random digit dial surveys who might be inappropriately identified as likely voters, using information generally gathered in commercial voter opinion studies (as might be used to influence a legislator or public opinion generally). The second and third research questions support this primary question and ground it in practical matters of public policy: Are misidentified likely voters different than true likely voters (as suggested by the literature)? Does it matter – do the two groups actually differ on policy or political issues?

Amount of Over Reporting

The validation study, on which this thesis is based, indicated that significant numbers of respondents over reported prior voting habits. This validation study consisted of comparing respondents' recollection of voting in particular election with official California records, information which was included in the sample, as discussed in Chapter 3. Table 3 shows the percentage of respondents who over reported for each of the four test elections, and Table 4 shows the distribution of over reporting among respondents. Notably, only 29.9% of respondents did not over report for any test elections.

TABLE 3: RESPONDENT OVER REPORTING BY ELECTION

Election	*% Respondents*
November 1992	44.3
November 1994	45.3
November 1996	43.3
November 1998	27.8

TABLE 4: DISTRIBUTION OF RESPONDENT OVER REPORTING

OverRepScore	*% Respondents*
0 Elections	29.9
1 Election	20.3
2 Elections	19.5
3 Elections	20.3
4 Elections	10.1

Research Question #1: Feasibility of Predicting Vote Overreporters

This section presents results of a series of regression tests of common explanatory variables to explore their effect on two variations of a dependent variable indicating vote over reporting by a respondent. A dichotomous version of the dependent variable indicates *if* a respondent over reported prior voting behavior, and a scalar version of the dependent variable indicates *how much* a respondent over reported prior voting behavior. The following sections present descriptions and results of the regression analyses and discussion of the goodness of fit of the various models explored.

Evaluating the Regression Results

Because this analysis involves exploration of several variations of the basic regression model, I expect interpretation of the results to present significant challenges. The different dependent variables and functional forms are likely sensitive to different variables, and possibly even different aspects of the same variables. Indeed, this variation in sensitivity is why I chose this exploratory approach. Consequently, I am not looking for a simple, direct answer to the question of the feasibility of predicting over reporters. Rather, my purpose is to see what indications of feasibility arise from this exploration.

Some of these indications are relatively straightforward. One surface indication is to see how the regression results compare to the expected impact of variables described in Chapter 3. In addition, I would expect those variables with a significant effect to have a strong effect on whether a respondent is likely to over report his or her voting history. Perhaps most importantly, I would expect a strong result from the tests for the overall goodness of fit of the model; that is, I would expect the model as a whole to predict which respondents would over report more correctly than would random assignment.

Other indications may be more subtle. For example, if prediction with this data were feasible, I would expect to see some consistency in which variables are significant across the regression variations, and particularly in the direction of effect. That said, one of the values of exploring different variations of the functional form of the model is to look for differences in how different variables affect the outcome. In this light, some

variables may show differences in the strength of their effect in different variations of the model, though I would still expect consistency in direction.

The following subsections present the results of the variations of the regression model in light of these indications. The broader aspects of these results are discussed at the conclusion of this chapter.

Results of Regression Analyses

This section includes the results of six regression models exploring the effects of variations of the explanatory variables described in Chapter 3 on dichotomous and scalar versions of the dependent variable. Each subsection describes the regression models used and treatment of variables, and includes a table of results. The first subsection describes comparison of logistic regression and linear (OLS) regression of a dichotomous dependent variable. The second subsection describes comparison of several functional forms of linear regression and certain treatment of explanatory variables. The third subsection describes potential sources of error in the regression models.

Logistic regression with dichotomous dependent variable.

The results from the logistic regression are somewhat mixed and do not clearly support the feasibility of predicting over reporters with this model. Three variables showed indications of effects in the expected direction, significant at a 95% level of confidence. However, for two of the three, only one category showed a significant effect; not enough to draw firm conclusions from. Additionally, the strength of effect

was relatively weak, though all three showed clear improvement in predictive power over the OLS linear regression.

In contrast, one variable, having an unlisted phone number, was significant though I expected it to have no effect, and had included it in the model as a control. It too had a relatively weak effect and the logistic regression model showed improvement over the OLS model.

Table 5 contains detailed results of both a logistic regression and a linear (OLS) regression for comparison. Because logistic regression requires " a reasonable representation of both alternative choices," (Studenmund, 2006, p. 456), I reduced the sample to provide balance prior to performing the regressions. The original sample of 800 interviews included 561 respondents who over reported prior voting behavior and 239 who did not. To balance the sample, I randomly selected 240 respondents from among those who over reported, providing a balanced sample of 240 over reporters and 239 non-overreporters. I ran logistic regressions on both the reduced sample and the full sample, to test the effects of balancing. Additionally, I ran a linear (OLS) regression on the full sample, to compare against the logistic model.

Interpretation of results is slightly different for the two methods. Beta coefficients for the linear OLS method indicate the number of units the dependent variable will change for a one-unit change in a given independent variable. Interpretation of the Exp(B) coefficients in logistic regression yields the percentage change in the dependent variable for a one-unit change in a given independent variable. Interestingly, because the dependent variable is dichotomous, the Beta coefficient from

OLS has the effect of indicating a percentage change; specifically the percentage of cases expected to change "state" due to a one-unit change in the independent variable. This renders the two statistics a bit more comparable, though the linear method is still subject to boundary problems, and does not accurately reflect the change in slope of the estimation curve.

TABLE 5: COMPARISON OF OLS & LOGISTIC REGRESSION RESULTS

	Full Sample			*Reduced Sample*		
Variable Label	***OLS Beta (Sig.)***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***VIF (OLS)***
(Constant)		2.697 (0.042)	169.7	0.621 (0.443)	-37.9	
Independent: Socio-Economic Status						
Employment: Student	-0.011 (0.779)	0.713 (0.467)	-28.7	0.607 (0.359)	-39.3	1.383
Employment: Homemaker	-0.025 (0.508)	0.873 (0.728)	-12.7	1.142 (0.767)	14.2	1.222
Employment: Retired	-0.044 (0.399)	0.621 (0.097)	-37.9	0.433* (0.034)	-56.7	2.376
Employment: Part Time	-0.006 (0.863)	0.739 (0.369)	-26.1	0.878 (0.738)	-12.2	1.17
Employment: Full Time (*Category Omitted*)						
Employment: Unemployed	-0.019 (0.584)	0.617 (0.344)	-38.3	0.425 (0.224)	-57.5	1.1
Education: Grades 1 to 8	-0.025 (0.485)	0.897 (0.897)	-10.3	1.416 (0.739)	41.6	1.123
Education: Grades 9 to 11	-0.027 (0.458)	0.716 (0.586)	-28.4	0.203 (0.132)	-79.7	1.187
Education: HSGrad	-0.021 (0.597)	1.214 (0.484)	21.4	1.237 (0.541)	23.7	1.4
Education: Some College	-0.036 (0.385)	0.96 (0.86)	-4	0.868 (0.621)	-13.2	1.489
Education: College Grad (*Category Omitted*)						
Education: Post Grad	0 (0.993)	0.854 (0.541)	-14.6	0.665 (0.205)	-33.5	1.376
Home_Owner	-0.028 (0.487)	0.995 (0.983)	-0.5	1.101 (0.756)	10.1	1.408
Years in Contra Costa	-0.16* (0.001)	0.977* (0.001)	-2.3	0.974* (0.001)	-2.6	1.402

TABLE 5: COMPARISON OF OLS & LOGISTIC REGRESSION RESULTS

	Full Sample			*Reduced Sample*		
Variable Label	***OLS Beta (Sig.)***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***VIF (OLS)***
Income: 10K or Less	-0.086 (0.022)	0.598 (0.323)	-40.2	0.63 (0.473)	-37	1.238
Income: 10,001 - 20,000	-0.032 (0.396)	0.483 (0.129)	-51.7	0.605 (0.413)	-39.5	1.266
Income: 20,001 - 30,000	0.003 (0.946)	1.516 (0.324)	51.6	0.967 (0.953)	-3.3	1.346
Income: 30,001 - 40,000	-0.008 (0.829)	0.721 (0.312)	-27.9	0.988 (0.976)	-1.2	1.278
Income: 40,001 - 50,000	-0.033 (0.373)	1.026 (0.946)	2.6	0.8 (0.636)	-20	1.223
Income: 50,001 - 60,000	0.022 (0.553)	1.062 (0.852)	6.2	0.665 (0.327)	-33.5	1.24
Income: 60,001 - 70,000	-0.011 (0.768)	0.816 (0.526)	-18.4	0.692 (0.35)	-30.8	1.21
Income: 70,001 - 80,000	0.023 (0.539)	1.652 (0.226)	65.2	1.645 (0.306)	64.5	1.189
Income: 80,001 - 100,000	-0.046 (0.22)	0.769 (0.406)	-23.1	0.904 (0.788)	-9.6	1.224
Income: Over 100,000 (*Category Omitted*)						
Independent: Politics						
Party: Rep	-0.043 (0.312)	0.882 (0.569)	-11.8	1.188 (0.554)	18.8	1.558
Party: Dem (*Category Omitted*)						
Party: DTS	0.027 (0.468)	1.457 (0.296)	45.7	1.787 (0.161)	78.7	1.185
Party: Other	0 (0.993)	0.815 (0.57)	-18.5	0.595 (0.27)	-40.5	1.164
Political Philosophy (Conservative – Liberal)	0.055 (0.156)	1.104 (0.282)	10.4	1.22 (0.105)	22	1.333
Independent: Personal Demographics						
Age: 18 – 29	-0.058 (0.199)	1.001 (0.997)	0.1	1.733 (0.189)	73.3	1.776
Age: 30 – 39	0.097* (0.021)	2.016* (0.02)	101.6	3.564* (0.001)	256.4	1.535
Age: 40 – 49 (*Category Omitted*)						
Age: 50 – 64	0.005 (0.916)	1.049 (0.852)	4.9	1.37 (0.327)	37	1.63
Age: 65 and Over	-0.091 (0.109)	0.697 (0.294)	-30.3	1.358 (0.503)	35.8	2.787

TABLE 5: COMPARISON OF OLS & LOGISTIC REGRESSION RESULTS

	Full Sample			*Reduced Sample*		
Variable Label	***OLS Beta (Sig.)***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***Logistic Exp(B) (Sig.)***	***Logit Interp. †***	***VIF (OLS)***
Gender: Female	0.057 (0.121)	1.569* (0.017)	56.9	1.454 (0.109)	45.4	1.188
Ethnicity: White/Cauc. (*Category Omitted*)						
Ethnicity: Black/African American	0.087* (0.02)	2.348 (0.067)	134.8	2.273 (0.143)	127.3	1.209
Ethnicity: Hispanic/Latino	0.044 (0.224)	1.35 (0.442)	35	1.78 (0.207)	78	1.13
Ethnicity: Asian	-0.068 (0.056)	0.551 (0.191)	-44.9	0.571 (0.309)	-42.9	1.1
Ethnicity: Other	0.009 (0.79)	1.346 (0.546)	34.6	1.451 (0.528)	45.1	1.076
Phone_Unlisted	0.114* (0.002)	1.71* (0.016)	71	1.788* (0.027)	78.8	1.127

* Significant at 95% level of confidence; † Probability = (Exp(B)-1) * 100

Linear regression with scalar dependent variable.

Results of the linear regression models, like those of the logistic models, are also mixed, and lack clear indications of support for the feasibility of predicting over reporters. Four variables showed consistent effect across functional forms, though only two, length of residence (Years in Contra Costa) and income, were variables I expected to have significance. The other two, unlisted phone numbers and ethnicity, were controls that I did not expect to effect respondents' propensity to over report. Strangely, the direction of effect for Years in Contra Costa is negative in the linear models, and positive in the logistic models. This is perhaps the best example of the mixed nature of the results, suggesting that the variable is significant, but that the models may not be

adequate to understand its effect. Other results discussed below seem to support this idea.

Table 6 includes detailed results of four linear regressions for comparison. Unlike the previous comparison, all of these models used the full sample of 800 interviews. The first regression is the same OLS model used in the OLS/Logistic comparison, using dummy variables for all values except length of residence (Years in Contra Costa). The other three regressions used scalar equivalents for Age, Education, and Income to allow for a comparison of functional forms, including OLS (Lin-Lin), Log-Linear, and Log-Log. I used linear regression for all four, replacing variables with calculated logs as appropriate.

Similar to the previous comparison of logistic regression and OLS, the effects of all of the significant variables are weak. As I expected, there are variations in which functional form is most sensitive to a particular explanatory variable, but the weakness of their effect suggests that this is again insufficient to draw conclusions about. Similarly, the scalar version of the variable for income show indications of significance in the linear models, while the categorical version in the logistic, and OLS comparison, models does not. This appears to be consistent with the idea of a specification problem mentioned above.

Another interesting pattern is that several categories related to ethnicity showed significance in one or two variations each (none were significant in the logistic model). The remainder of the results is included in the table.

TABLE 6: COMPARISON OF LINEAR REGRESSION RESULTS

	Dummy Vars.		*Scalar Variables*					
Functional Form:	*Lin-Lin*		*Lin-Lin*		*Log-Lin*		*Log-Log*	
Variable Label	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***
Constant			(B) 1.044 (0.003)		(B) 0.246 (0.144)		(B) 0.258 (0.178)	
Independent: Socio-Economic Status								
Employment: Student	-0.011 (0.779)	1.383	-0.4 (0.13)	1.196	-0.046 (0.356)	1.130	-0.024 (0.626)	1.158
Employment: Homemaker	-0.025 (0.508)	1.222	0.011 (0.96)	1.194	-0.003 (0.956)	1.215	-0.01 (0.852)	1.219
Employment: Retired	-0.044 (0.399)	2.376	-0.131 (0.464)	1.970	0.06 (0.334)	1.767	0.05 (0.386)	1.527
Employment: Part Time	-0.006 (0.863)	1.17	-0.044 (0.826)	1.160	0.022 (0.668)	1.165	0.026 (0.607)	1.169
Employment: Full Time (*Category Omitted*)								
Employment: Unemployed	-0.019 (0.584)	1.1	-0.067 (0.839)	1.079	-0.012 (0.812)	1.094	-0.016 (0.745)	1.090
Education (Scalar)			0.063 (0.244)	1.326	0.099 (0.072)	1.369		
Log (Education - Scalar)							0.08 (0.138)	1.335
Education: Grades 1 to 8	-0.025 (0.485)	1.123						
Education: Grades 9 to 11	-0.027 (0.458)	1.187						
Education: HSGrad	-0.021 (0.597)	1.4						
Education: Some College	-0.036 (0.385)	1.489						
Education: College Grad (*Category Omitted*)								
Education: Post Grad	0 (0.993)	1.376						
Home Owner	-0.028 (0.487)	1.408	-0.17 (0.252)	1.522	-0.054 (0.345)	1.515	-0.065 (0.251)	1.498
Years in Contra Costa	-0.16* (0.001)	1.402	-0.012* (0.002)	1.424	-0.088 (0.102)	1.317		
Log (Yrs in Contra Costa)							-0.145* (0.005)	1.217
Income (Scalar)			0.072* (0.005)	1.742	0.115* (0.07)	1.821		
Log (Income - Scalar)							0.135* (0.029)	1.781

TABLE 6: COMPARISON OF LINEAR REGRESSION RESULTS

	Dummy Vars.		*Scalar Variables*					
Functional Form:	*Lin-Lin*		*Lin-Lin*		*Log-Lin*		*Log-Log*	
Variable Label	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***
Income: 10K or Less	-0.086 (0.022)	1.238						
Income: 10,001 - 20,000	-0.032 (0.396)	1.266						
Income: 20,001 - 30,000	0.003 (0.946)	1.346						
Income: 30,001 - 40,000	-0.008 (0.829)	1.278						
Income: 40,001 - 50,000	-0.033 (0.373)	1.223						
Income: 50,001 - 60,000	0.022 (0.553)	1.24						
Income: 60,001 - 70,000	-0.011 (0.768)	1.21						
Income: 70,001 - 80,000	0.023 (0.539)	1.189						
Income: 80,001 - 100,000	-0.046 (0.22)	1.224						
Income: Over 100,000 (*Category Omitted*)								
Independent: Politics								
Party: Rep	-0.043 (0.312)	1.558	-0.115 (0.402)	1.509	-0.039 (0.496)	1.482	-0.04 (0.481)	1.479
Party: Dem (*Category Omitted*)								
Party: DTS	0.027 (0.468)	1.185	-0.005 (0.98)	1.161	0.01 (0.84)	1.167	0.009 (0.851)	1.165
Party: Other	0 (0.993)	1.164	-0.181 (0.401)	1.137	0.018 (0.716)	1.121	0.026 (0.599)	1.121
Political Philosophy (Conservative – Liberal)	0.055 (0.156)	1.333	0.113 (0.045)	1.314	0.09 (0.09)	1.294	0.082 (0.122)	1.297
Independent: Personal Demographics								
Age (Scalar)			-0.067 (0.263)	2.160	0.042 (0.506)	1.865		
Log (Age - Scalar)							0.09 (0.126)	1.606
Age: 18 - 29	-0.058 (0.199)	1.776						
Age: 30 - 39	0.097* (0.021)	1.535						
Age: 40 - 49 (*Category Omitted*)								

TABLE 6: COMPARISON OF LINEAR REGRESSION RESULTS

	Dummy Vars.		*Scalar Variables*					
Functional Form:	*Lin-Lin*		*Lin-Lin*		*Log-Lin*		*Log-Log*	
Variable Label	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***	***Beta (Sig.)***	***VIF***
Age: 50 - 64	0.005 (0.916)	1.63						
Age: 65 and Over	-0.091 (0.109)	2.787						
Gender: Female	0.057 (0.121)	1.188	0.137 (0.227)	1.201	-0.03 (0.567)	1.212	-0.026 (0.612)	1.211
Ethnicity: White/Cauc. (*Category Omitted*)								
Ethnicity: Black/African American	0.087* (0.02)	1.209	0.497* (0.02)	1.139	0.076 (0.133)	1.169	0.081 (0.109)	1.185
Ethnicity: Hispanic/Latino	0.044 (0.224)	1.13	0.348 (0.099)	1.113	0.098* (0.049)	1.123	0.099* (0.046)	1.129
Ethnicity: Asian	-0.068 (0.056)	1.1	-0.605* (0.019)	1.098	-0.083 (0.087)	1.079	-0.082 (0.09)	1.077
Ethnicity: Other	0.009 (0.79)	1.076	0.012 (0.964)	1.061	0.01 (0.829)	1.082	0.015 (0.762)	1.078
Phone: Unlisted	0.114* (0.002)	1.127	0.458* (0.001)	1.104	0.117* (0.018)	1.110	0.114* (0.019)	1.101
* Significant at 95% level of confidence								

Sources of Potential Error

This section discusses sources of potential error in the regression analyses described above, particularly including multicolinearity and heteroskedasticity.

Multicolinearity.

Multicolinearity refers to linear relationships between supposedly independent variables in a regression model. As part of each regression analysis, I tested for multicolinearity and found no indications of its presence.

As discussed in Chapter 3, the modal category from each variable group was omitted to reduce the potential for multicolinearity. Though omission of other

categories might be beneficial for purposed of analysis, for example because they may tell a better story of comparison, omitting other categories during test runs created significant multicolinearity problems in several variable groups.

For diagnosing multicolinearity, the SPSS software provides a Variance Inflation Factor (VIF) statistic for each independent variable in a regression model. This statistic reflects "the extent to which a given explanatory variable can be explained by all the other explanatory variables in the equation" (Studenmund, 2006, p. 258). Normally, multicolinearity is considered to be present and problematic if the VIF statistic is greater than 5. All VIF statistics were less than 5, and are included in the table with the results for each regression model.

Heteroskedasticity.

Heteroskedasticity refers to changes in the distribution of stochastic error due to changes in the scale of a critical variable. As part of each regression analysis, I tested for heteroskedasticity. The presence of heteroskedasticity was indicated in one of the two logistic regression models. Tests for heteroskedasticity on the other five regression models were negative. The logistic regression testing positive was run on the full sample of 800 interviews. Distinguishing characteristics of this regression run are discussed below.

Studenmund (2006, p. 346) points out that "in general, heteroskedasticity is more likely to take place in cross-sectional models than in time-series models." Heteroskedasticity represents either an effect related to the scale of an explanatory variable (e.g. big cities have characteristics small cities lack), or an error in specification

causing changes in the distribution of the stochastic error term (e.g. information about city size was omitted), which could appear as though there were a scale-related effect.

Possible sources of heteroskedasticity in this study include one interval variable (Years in Contra Costa), and several variables that are treated as scalar for purposes of exploring the effect of functional form on the linear regression model (Age, Education, and Income). I used two different tests for heteroskedasticity: the Park test, and the White test. The remainder of this section discusses these tests and results are presented in Table 7.

The logistic regression models and the first variation of the linear regression model included dummy variables for values of all variables except respondents' length of residence (Years in Contra Costa). I used the Park test to test for heteroskedasticity in these regression models. This test regresses the log of the squared residuals from a regression with potentially heteroskedastic properties against the log of the variable likely causing the scale effect. The test produces a coefficient which can be tested using a basic t-test to determine the presence of heteroskedasticity. Corrections can then be made as necessary by weighting the data along that variable.

As indicated above, the results of all of the Park tests were negative, with one exception. The Park test for the logistic regression run on the full sample of 800 interviews indicated the presence of heteroskedasticity. As discussed earlier, however, logistic regression works best when the number of cases is relatively balanced, and the full sample is not evenly balanced. In this light, it is perhaps not surprising that test parameters are outside of norms for this regression run.

Three of the linear regression models included scalar variations of variables for Age, Education, and Income, as well as including Years in Contra Costa. I used to White Test to test for heteroskedasticity in these regression models. Similar to the Park test, this test includes factors for potential interactions between suspicious variables. Specifically,

the test regresses the log of the squared residuals from a regression with potentially heteroskedastic properties against each suspicious variable, the square of each suspicious variable, and the products of each suspicious variable with every other suspicious variable. The test produces a test statistic, NR^2 (where N is the sample size and R^2 is the unadjusted R^2 coefficient of determination of the test regression), which can be compared against the critical value from standard chi square tables, with degrees of freedom equal the number of explanatory variables in the test regression. The results of all of the White tests were also negative.

TABLE 7: RESULTS OF PARK AND WHITE TESTS

Model	*Test Statistic*	*Critical Value or Sig. (1%, two tailed)*
Logistic (reduced sample)	$t = 1.476$	(.141)
Logistic (full sample)	$t = 6.022$	(.001)
Linear (all dummies)	$t = 0.000$	(1.000)
Linear (scale vars)	$NR^2 = 10.523$	$\chi^2 = 29.1$
Log-Lin (scale vars)	$NR^2 = 19.536$	$\chi^2 = 29.1$
Log-Log (scale vars)	$NR^2 = 23.976$	$\chi^2 = 29.1$

Other sources of potential error.

Another source of potential error in these regression analyses is specification of variables. For logistic regression it is useful to have as many variables as possible converted into dummy variables for each response value. For linear regression, however, more "resolution" for looking into the data is helpful, such as is provided by interval variables. Unfortunately, most of the data in this survey was captured categorically, mainly because of the difficulty in eliciting detailed and accurate information from respondents.

An additional element of potential specification error in surveys such as this is in definition of categories. For example, age was captured by asking in what year the respondent was born (respondents are more likely to provide accurate birth years than accurate ages), and, the data was coded into roughly ten year age categories: 18-29, 30-39, 40-49, 50-64, and 65 or over; relatively common categories. Obviously, this is arbitrary and does not account for generational cohorts or other issues more subtle than the general age of the respondent.

Goodness of Fit and Comparison of Functional Forms

Overall, the models explored were not very good at predicting the propensity of a respondent to over report prior voting history. The logistic model, when run on the balanced sample produced the best result, improving prediction to 21 percentage points from the likelihood of 50% accuracy through random assignment. However, this is a circular result as that sample was artificially balanced to improve the regression results.

The same model only improved prediction from 70.4% to 73.4 in the unbalanced full sample that was randomly selected from the target population.

The linear regression models also showed poor overall fit, which clearly worsened as the model moved to logarithmic forms. Interestingly, the two scalar variables that showed significant effect, Years in Contra Costa and Income, had their strongest effects in the log-log form of the regression. This seems consistent with the specification problem I discussed earlier.

Table 8 includes the results of goodness of fit tests for each model. In linear regression, goodness of fit is indicated by the R^2 statistic. The test statistic lies between zero and one, with a value near one representing a good fit (Studenmund, p. 50). Because logistic regression is non-linear, however, this statistic is not reliable. While there are several methods for estimating the goodness of fit of a logistic regression model, the simplest is the combined percentage of cases correctly predicted by the equation. In this case, that means the average of the percentage of correctly predicted over reporters and the percentage of correctly predicted non over reporters. The table includes the R^2 statistics for the linear regression models and mean percentage of correct predictions for the logistic regression model.

TABLE 8: GOODNESS OF FIT

Model	*R^2*	*Mean % Correct*
Logistic (reduced sample)		71.0
Logistic (full sample)		73.4
Linear (all dummies)	.159	
Lin-Lin (scale vars)	.143	
Log-Lin (scale vars)	.076	
Log-Log (scale vars)	.093	

Research Question #2: Are Vote Over-Reporters Different?

The literature suggests that vote over reporters are different demographically from non over reporters. As discussed in Chapter 2, Belli, Traugott, and Beckman (2001) reported differences in age, education, strength of political attitudes, and ethnicity. In this thesis I am focusing a little more tightly on respondents identified as likely voters. Based on the literature, if identified likely voters differ between over reporters and non over reporters, I would expect to find similar differences in the demographics. The results reported below show that over reporting likely voters are, in fact, significantly different from non over reporting likely voters in most demographic categories.

These results show differences between over reporting likely voters and non over reporting likely voters that are consistent with the results reported by Belli, Traugott, and Beckman. I found significant differences between the two groups in age and ethnicity at a 95% level of confidence. While I did not have a direct test for strength of political attitudes, I did find differences between the two groups on a test of political philosophy: a Likert scale where values from 1 to 5 represent very conservative, somewhat conservative, moderate, somewhat liberal, and very liberal. Though this is not a direct test of attitudinal strength, this structure does include a component indicating strength of attitude. In contrast, I did not find a difference between the two groups in education.

The remainder of this section includes discussion of the tests used to compare over reporting likely voters with non over reporting likely voters, and presentation of

the results of those tests. These tests used the original versions of each variable, instead of the recoded dummies used for many of the regression models. This allowed comparison of categorical (nominal) explanatory variables with the chi square statistic, and comparison of means for scalar variables, as described below.

Table 9 includes chi square and related statistics for the nominal variables. The significance statistic "Asymp. Sig. (2-sided)" represents a P-value for the probability that random sampling will produce the observed results, assuming the null hypothesis is correct that there is no relationship between the dependent variable and the selected explanatory variable (Pollock, 2009, p. 139). For example, if the null hypothesis is correct that there is no relationship between Likely Voters and Education in the sampled population, then random sampling will produce the observed data 34.0 percent of the time.

Both the Lambda and Cramer's V statistics also provide indications of the strength of relationship between the two variables, ranging between 0 (weak) and 1 (strong). Lambda "measures the *percentage of improvement* in guessing values on the dependent variable on the basis of knowledge of values on the independent variable" (Manheim, Rich, & Willnat, 2002), but can fail to detect a relationship when the mode of the dependent variable for each category of the independent variable is the same. Cramer's V provides a backup for such situations (Pollack, p. 146).

Table 10 includes comparison of means for two scalar variables, Years in Contra Costa (interval), and Political Philosophy (ordinal).

TABLE 9: CHI SQUARE RESULTS FOR RESEARCH QUESTION #2

Dependent Variable: Likely Voters (Over Reporting v. Non Over Reporting)					
Variable Label	***X^2***	***df***	***Asymp. Sig. (2-sided)***	***Lambda***	***Cramer's V***
Independent: Socio-Economic Status					
Employment	59.044*	6	.001	.228	.291
Education	6.802	6	.340	.015	.099
Home Owner	18.896*	2	.001	.019	.164
Household Income	19.214	11	.057	.099	.166
Independent: Politics					
Party	17.400*	5	.004	.099	.158
Independent: Personal Demographics					
Age	80.833*	5	.001	.225	.340
Gender	5.174*	1	.023	.019	.086
Ethnicity	12.145*	5	.033	.001	.132
Phone Unlisted	23.619*	3	.001	.056	.184
* Significant at 95% level of confidence					

TABLE 10: COMPARISON OF MEANS FOR RESEARCH QUESTION #2

Dependent Variable: Likely Voters (Over Reporting v. Non Over Reporting)				
Variable Label	***z***	***p***	***Mean (Over Reporting)***	***Mean (Non Over Reporting***
Independent: Socio-Economic Status				
Years in Contra Costa	7.714*	0.001	19.02	28.23
Independent: Politics				
Political Philosophy (Conservative – Liberal)	2.962*	0.003	3.0028	2.7564
* Significant at 95% level of confidence				

Research Question #3: Does it Matter?

This section examines results from the third research question exploring differences on policy or political issue areas between over reporting likely voters and non over reporting likely voters. Table 11 contains chi square statistics for these comparisons. These tests indicated significant differences between the two populations

in two areas: reaction to arguments regarding support for California's then recently established open primary, and support for a potential local school bond in light of various levels of a hypothetical property tax increase to pay for it.

TABLE 11: CHI SQUARE RESULTS FOR RESEARCH QUESTION #3

Dependent Variable: Likely Voters (Over Reporting v. Non Over Reporting)					
Variable Label	***X^2***	***Df***	***Asymp. Sig. (2-sided)***	***Lambda***	***Cramer's V***
Issue Area #1: Right Direction or Wrong Track					
Your Community	.101	2	.951	.001	.012
California	1.475	2	.478	.001	.046
The Country	1.275	2	.529	.001	.043
Issue Area #2: Open Primary					
Support Open Primary	17.158	4	.002	.102	.157
Change Open Primary	3.866	4	.424	.009	.074
Issue Area #3: Head to Head 2000					
Bush v. Gore	6.086	4	.193	.040	.093
Issue Area #4: Local School Bonds (potential annual property tax increase)					
$53	9.673	4	.046	.052	.118
$42	9.977	4	.041	.062	.119
$36	10.798	4	.029	.068	.124
$27	9.528	4	.049	.056	.117

On the issue of California's open primary, there are several significant aspects worthy of note. First, the chi square test indicates significant differences on the first question of support for the open primary as it stood. The chi square test indicates no significant difference however, when respondents were asked about changing the open primary after hearing additional arguments that the results would not be counted by the national Republican or Democratic parties.

Second, there were significant differences in how over reporting likely voters responded to the intermediate argument, compared to non over reporting likely voters. Table 12 shows the change in the so called "top 2 box" (respondents who either "support strongly" or "support somewhat" the issue in question). For both questions, a higher percentage of over reporting likely voters responded in the top 2 box. Additionally, the percentage change for over reporting likely voters was 16.5 percentage points greater, or about 60% larger than the swing for non over reporting likely voters.

TABLE 12: CHANGE IN SUPPORT (TOP 2 BOX) FOR OPEN PRIMARY

Question	***Over Reporting Likely Voters***	***Non Over Reporting Likely Voters***
Support open primary (top 2 box)	302 (80.5%)	218 (67.3%)
Change open primary (top 2 box)	239 (63.8%)	196 (60.5%)
Percent change in top 2 box	166 (44.3%)	90 (27.8%)
N	*375*	*324*

On the issue of the school bond, over reporting likely voters supported a potential local school bond measure at higher levels of potential property tax increases. Table 13 includes top 2 box frequencies and percentage of support for each level.

TABLE 13: SUPPORT (TOP 2 BOX) FOR POTENTIAL LOCAL SCHOOL BOND

Level of Potential Property Tax Increase	***Over Reporting Likely Voters***	***Non Over Reporting Likely Voters***
$53	268 (71.5%)	206 (63.6%)
$42	275 (73.3%)	153 (65.7%)
$36	281 (74.9%)	216 (52.6%)
$27	283 (75.5%)	231 (71.3%)
N	*375*	*324*

Summary of Findings

Several significant findings emerged from this analysis. To begin with, large numbers of respondents over reported prior voting behavior. Regarding the exploration of regression models, there was little support for the feasibility of predicting vote over reporters with the data and model I used. While a few variables were consistently significant across variations of the regression model, all significant variables had relatively weak effects on respondents' propensity to over report regardless of the regression variation. Additionally, there were some indications of problems with specification of the model as well as the structure of the data.

Comparison of over reporting likely voters and non over reporting likely voters found significant differences between the two groups in socio-economic status (employment status, home ownership, and length of residence), demographics (age, gender, ethnicity, and unlisted phone numbers), and in political attitudes (party registration and conservative – liberal political philosophy).

Additionally, over reporting likely voters and non over reporting likely voters differed on some matters of policy and politics. Over reporting likely voters were more supportive of the then recently passed open primary system, but changed their views more after hearing follow-up arguments. Over reporters were also willing to support a hypothetical local school bond measure at higher levels of potential property tax increases than non over reporting likely voters.

Further discussion of these findings and their implications, and conclusions support by this analysis, is included in Chapter 5.

Chapter 5

CONCLUSIONS AND IMPLICATIONS

This chapter includes discussion of the results presented in Chapter 4 and the conclusions they support, along with a review of the limitations of this analysis and suggestions for future research.

Discussion of Conclusions

This section includes conclusions related to each of the three research questions of this thesis and their implications, as well as more general issues suggested by the results.

Vote Over Reporting

Vote over reporting of prior voting behavior is an ongoing problem affecting a significant number of prospective respondents to voter opinion surveys. The finding that significant numbers of respondents over reported their voting history is consistent with the literature that vote over reporting is a historical and widespread issue. In that context this finding is unsurprising, yet it is critical. It serves as a threshold issue, establishing the relevance of the issues researched here. It also provides further documentation of the pattern of vote over reporting by survey respondents and opens the door to exploration of how related findings in the literature apply to these data.

If it is not surprising that many respondents over reported voting, what may be startling is how *many* respondents did so. The literature reviewed in Chapter 2 documents rates of over reporting ranging from 7.8% to 28% of respondents inaccurately reported voting in at least one election. In contrast, only 29.9% of

respondents in this study *did not* over report voting in at least one test election. Fully 30.4% of respondents over reported voting in at least 3 of the 4 test elections.

These findings are clearly different from the literature and raise questions about whether rates of vote over reporting have increased over time, whether these data represent an outlier population, or whether there was a systemic error in the preparation or collection of data. I consider the later unlikely because the survey used data collection methodologies specifically designed to prevent such systemic error: (1) a professional provider using their standard procedures prepared the sample, (2) the sample included voting history data for each potential respondent, (3) interviewers coded respondents' voting history directly on survey forms at the conclusion of each interview, and (4) computer data entry was segregated from the interview process and staff.

Research Question #1: Feasibility of Predicting Vote Overreporters

Using information generally gathered through surveys using random digit dial sampling methodologies to predict vote over reporting by prospective respondents is extremely difficult, and does not appear to be feasible. While the analysis presented in this thesis does not completely rule out the possibility of developing a usable model, it does not lend much support to the feasibility of doing so.

As presented in Chapter 4, the variables that showed significant effect were not consistent across variations of the regression models. However, even though they were not consistent in every specification, the variables that showed significant effect did tend to be the ones also identified as significant in the literature. For example, Belli,

Traugott, and Beckman (2001) found over reporters to be "predominantly non-white." My own logistic regression showed that ethnicity has no effect on *if* a respondent over reported, but the linear models showed indications that it did affect *how much* a respondent over reported. Even with the later, however, there are inconsistencies in which functional forms predict which ethnic groups more or less likely to over report: Asians less likely in one variation; Black/African Americans more likely in two variations; Hispanics more likely in logarithmic forms. All of this suggests that ethnicity can be a factor, but the models are not up to the task.

Another interesting finding relates to unlisted phone numbers. This was an important control to include because the key argument for use of random digit dial sampling is the risk that voters with unlisted phone numbers differ significantly from those with listed numbers. I found that respondents with unlisted phone numbers were in fact different from respondents with listed numbers. Ironically, however, respondents with unlisted phone numbers were *more* likely to over report prior voting behavior. This implies that respondents with unlisted phone numbers were actually less likely to vote, and that random digit dial surveys are likely to over represent their opinions. This would seem to undermine the argument for use of RDD sampling.

Research Question #2: Are Vote Over Reporters Different?

Second, misidentified "likely voters" are different in demographics and political outlook than correctly identified likely voters. The second research question focuses the analysis on identified likely voters, asking if those who were mis-identified because they over reported prior voting behavior differ from those who were correctly

identified. The data clearly show the two groups are different in nearly all demographic categories. This is consistent with findings in the literature that over reporters and likely voters represent two different populations. A key divergence from the literature is that the two groups did not differ on level of education (nor was education significant in the regression analysis), but this may be due to relatively high socioeconomic status for the target population as a whole.

Differences between over reporting likely voters and non over reporting likely voters included party registration and "conservative – liberal" political philosophy. This indicates that the two groups differ not only demographically, but attitudinally as well; suggesting that there are likely differences in political behavior and documenting a need to explore differences in policy or political preferences.

Research Question #3: Does it Matter – Do Vote Overreporting Likely Voters Differ on Policy Preferences?

Misidentified likely voters have significantly different preferences on issues of public policy, and surveys including them in the sampling frame risk biased results. As with research question #2, the third research question also resulted in clear evidence of difference between over reporting likely voters and non over reporting likely voters, this time regarding preferences on issue areas related to policy or politics. It is interesting that this difference did not manifest itself on the superficial "horserace" type questions so often publicized, such as who respondents prefer for President, or if they think things are on the right track or heading in the wrong direction. Rather, the differences became

evident on more substantive issues such as how much should we pay for our schools or how should California conduct elections.

As discussed in Chapter 2, Baldassare (2006) noted that non-voters "favor ballot measures that would spend more on programs to help the poor" while likely voters are "ambivalent and divided along party lines on ballot measures that would spend more on the poor." This is clearly consistent with the finding here that over reporting likely voters supported the hypothetical local school bond measure at higher levels of potential property tax increases than did non over reporting likely voters.

The test of arguments regarding the open primary also yielded interesting results. What is perhaps most interesting about this test, though, is that while the two populations started out with significantly different preferences about the open primary, they ended up fairly close together after hearing additional arguments. This does, however, lend itself to more than one possible interpretation. It could indicate that over reporters are less informed, and maybe not so different after all once knowledge levels are equalized. It could indicate that over reporters have less-firmly held preferences and are more susceptible to argument. A third interpretation may be more on point given the nature of the follow-up argument: that delegates selected through an open primary would not be counted by the national Republican and Democratic parties.

Good Citizens and Social Desirability Bias

Overall, the results of this study are consistent with the literature on social desirability bias, though not in the sense that I expected. Largely framed by the literature on the role of memory, my expectation was that any social desirability bias

would be an expression of established members of the community presenting the persona they felt they should have. This would have generated positive effects for factors such as age, length of residence, income, and education. In fact, the logistic regression did show a positive effect for length of residence (Years in Contra Costa) and the linear regression showed a positive effect for income, in two of the variations. I did not expect any effect from gender or ethnicity.

However, the remaining results were not consistent with my expectations. Length of residence was negative in the linear models; ethnicity showed positive effects for Black/African American and Hispanic/Latino, and negative for Asians; and the age category for 30 – 39 years old showed indications of a positive effect. The age category is interesting because it is the category just younger than the modal category (which was omitted).

This pattern of results suggests that the theory of social desirability bias in operation here was an expression of a desire to fit in not based in the self-perception of established members of the community, but as an expression of those who seek to become part of that establishment. In this sense, the social desirability pressure is not about behaving "as someone in my position should," but about behaving "as someone I want to be would act." The willingness to support the hypothetical local school bond at higher levels of potential property taxes would be consistent with this, if over reporting respondents' aspirational target includes a sense of civic mindedness.

Limitations of the Analysis

Like all research, this study is subject to several limitations provide context for interpretation of results. Since this thesis is largely about some of the limitations of certain types of survey research, clarity about the limitations of this study are perhaps even more important than usual.

Sampling Frame

As discussed in Chapter 3, the sampling frame for this survey was registered voters in Contra Costa County, California, and the survey was conducted in 1998. Extrapolating the results of any survey, either geographically or temporally, beyond its sampling frame carries significant risk. For example, other areas may have different social norms, or the social norms of the target population may change over time. It's often said that modern surveys are generally accurate, the problem is knowing who they are accurate about.

The target population for this survey was chosen because it closely reflected certain demographic characteristics of California as a whole, notably party registration, gender balance, and age distribution. This similarity opened the door to extrapolating the results to the broader population of registered voters in the state. Some of the survey results, however, indicate that other aspects of the demographics of the target population do not align as closely with those of the state. The modal category for household income, for example, was over $100,000. Similarly, the modal category for level of education was college graduate.

While income and education were not significantly different between over reporting likely voters and non over reporting likely voters, they have been identified in the literature and suggest that extrapolation of the results beyond the target population should be done carefully.

Cell Phones

Another limitation of this study is the advance of technology and related changes in behavior. When this survey was conducted in 1998, the use of cell phones was not nearly as widespread as it is today. The growing number of people who now use only cell phones has created significant problems for the survey industry. Voter file sampling relies on a combination of the phone numbers included on voter registration cards and matching voter registration information with phone numbers acquired through other sources, such as public records, to contact potential respondents. The growing use of cell phones, combined with growing concern about privacy, raise questions about the long term ability to generate sufficient phone numbers of registered voters for a valid sample.

While the growing use of cell phones is not a limitation on a survey already conducted, it does raise the possibility that attitudes have changed along with technology – which does create a limitation on the applicability of the results presented here to current populations.

Structure of Data

Most of the data used for this study were categorical in nature, as was appropriate for the type of data collected. However, the linear regression models using

scalar versions of several variables exhibited increased sensitivity to one of those variables (income) as well as to other variables in the model. This suggests that the model could be improved by collecting more specific data. This is not to say that doing so would be easy, as, for example, respondents particularly do not like sharing information about their income, but these results suggest that this is worth exploring. Based on these results, then, this should be considered a limitation of this analysis.

Suggestions for Future Research

The results and limitations of this study suggest several areas for future research. First, there appears to be a gap in the literature regarding differences between academic voter research and commercial voter research, and the methods commonly used by commercial researchers. The research team developed the questions used for this survey based on personal experience. Expanding the literature in this area would be beneficial in its own right, and would help provide a base from which future researchers can work.

The lack of support found for the feasibility of predicting over reporting using data commonly collected through random digit dial voter surveys supports the need for research into alternative screening techniques. In reviewing the literature, I did find that a few researchers are working on this issue. The findings that over reporters are different and have different policy preferences than true likely voters highlight the importance of methodologies that accurately sample a target population. If it is not feasible to do this with existing random digit dial sampling methods, and researchers choose not to use voter file sampling, then other methods of screening for RDD surveys must be developed.

Finally, though the results here did not provide much support for the feasibility of predicting over reporting using data generally collected through RDD surveys, the variables that were significant were generally in line with the literature, and there were indications that there are problems with this analysis which could be addressed in future research. Perhaps better data could be acquired or elements could be identified elements that might be included in a new methodology.

The Bottom Line

Taken together, the results of this study support four clear conclusions. First, over reporting of prior voting behavior is an ongoing problem affecting a significant number of prospective respondents to voter opinion surveys. Second, misidentified "likely voters" are different in demographics and political outlook than correctly identified likely voters. Third, misidentified likely voters have significantly different preferences on issues of public policy than correctly identified likely voters, and surveys including them in the sampling frame risk biased results. Finally, using information generally gathered through surveys using random digit dial sampling methodologies to predict vote over reporting by prospective respondents is extremely difficult, and does not appear to be feasible.

These conclusions tell a clear story, and it is a cautionary tale for both producers and consumers of voter opinion data. It's been said that decisions are made by those who show up. Likely voter surveys try to predict who that will be come election day. But methodology matters, and when surveys influence public policy and legislative decisions, it is imperative those surveys be accurate about whose attitudes and

preferences they present. Random digit dial sampling carries clear risk of bias toward those who want to be, rather than those who actually are “good citizens” (at least in terms of voting participation). As a result, RDD surveys are likely not to reflect the attitudes of true likely voters, and consumers of such surveys risk making public policy and law with bad information.

APPENDIX A

Correlation Matrix

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	EMPLOYMENT STATUS	Employment: Student	Employment: Homemaker	Employment: Retired
EMPLOYMENT STATUS	1	-.495**	-.431**	-.538**
Employment: Student	-.495**	1	-0.064	-.128**
Employment: Homemaker	-.431**	-0.064	1	-.155**
Employment: Retired	-.538**	-.128**	-.155**	1
Employment: Part Time	.072*	-.070*	-.085*	-.172**
Employment: Full Time	.874**	-.241**	-.290**	-.587**
Employment: Unemployed	-0.062	-0.039	-0.048	-.096**
EDUCATION	.170**	-0.061	0.02	-.158**
Education: Grades 1 to 8	-.088*	-0.024	0.017	.135**
Education: Grades 9 to 11	-0.068	-0.037	0.018	.098**
Education: HSGrad	-0.066	0.037	-.071*	.099**
Education: Some College	-0.053	.083*	0.007	-0.043
Education: College Grad	0.045	-0.024	0.067	-0.063
Education: Post Grad	.119**	-.079*	-0.031	-0.046
Home_Owner	0.025	-.249**	0.063	.147**
Yrs in Contra Costa	-.147**	-.088*	-.123**	.378**
INCOME	.185**	-.088*	.150**	-.282**
Income: 10K or Less	-.134**	.144**	0	0.037
Income: 10,001 - 20,000	-.110**	0.048	-0.028	.111**
Income: 20,001 - 30,000	-0.025	-0.062	-0.037	.114**
Income: 30,001 - 40,000	-0.018	0.004	-0.057	.079*

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	EMPLOYMENT STATUS	Employment: Student	Employment: Homemaker	Employment: Retired
Income: 40,001 - 50,000	0.024	0.001	-0.06	-0.001
Income: 50,001 - 60,000	0.057	0.002	-0.058	-0.016
Income: 60,001 - 70,000	.072*	-0.013	-0.037	-0.065
Income: 70,001 - 80,000	.097**	-0.018	0	-.108**
Income: 80,001 - 100,000	0.043	-0.013	.081*	-.096**
Income: Over 100,000	0.06	-0.033	.114**	-.146**
PARTY	0.025	0.008	-0.025	-0.024
Party: Dem	-0.002	-0.02	0.008	0.022
Party: Rep	-0.023	-.079*	0.028	0.058
Party: DTS	0.06	.084*	-0.039	-.110**
Party: Other	-0.022	.089*	-0.022	-0.023
Political Philosophy	0.041	.090*	-0.032	-.105**

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Employment: Part Time	Employment: Full Time	Employment: Unemployed	EDUCATION	Education: Grades 1 to 8
EMPLOYMENT STATUS	.072*	.874**	-0.062	.170**	-.088*
Employment: Student	-.070*	-.241**	-0.039	-0.061	-0.024
Employment: Homemaker	-.085*	-.290**	-0.048	0.02	0.017
Employment: Retired	-.172**	-.587**	-.096**	-.158**	.135**
Employment: Part Time	1	-.322**	-0.053	0.001	-0.033
Employment: Full Time	-.322**	1	-.180**	.188**	-.088*
Employment: Unemployed	-0.053	-.180**	1	-.110**	-0.018
EDUCATION	0.001	.188**	-.110**	1	-.332**
Education: Grades 1 to 8	-0.033	-.088*	-0.018	-.332**	1
Education: Grades 9 to 11	-0.049	-.072*	0.068	-.353**	-0.017
Education: HSGrad	-0.017	-.075*	0.065	-.580**	-0.047
Education: Some College	.101**	-.083*	0.057	-.255**	-0.068
Education: College Grad	-0.052	.085*	-.071*	.347**	-.074*
Education: Post Grad	-0.008	.116**	-0.062	.666**	-0.051
Home_Owner	-0.006	-0.026	-0.062	0.048	0.032
Yrs in Contra Costa	-0.055	-.194**	0.004	-.178**	.083*
INCOME	0.005	.209**	-0.056	.397**	-.086*
Income: 10K or Less	0.014	-.118**	0.049	-.119**	-0.02
Income: 10,001 - 20,000	0.012	-.123**	0.047	-.174**	.106**
Income: 20,001 - 30,000	-0.013	-0.048	0.012	-.172**	0.065
Income: 30,001 - 40,000	-0.024	-0.015	-0.03	-0.042	0.006

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Employment: Part Time	Employment: Full Time	Employment: Unemployed	EDUCATION	Education: Grades 1 to 8
Income: 40,001 - 50,000	0.033	0.001	0.037	-0.027	-0.03
Income: 50,001 - 60,000	-0.026	0.069	-0.031	-0.048	0.005
Income: 60,001 - 70,000	0.057	0.05	-0.003	0.018	-0.034
Income: 70,001 - 80,000	0.003	.105**	-0.018	0.062	-0.029
Income: 80,001 - 100,000	-0.005	0.067	-0.055	.143**	0.007
Income: Over 100,000	0.007	.074*	0.011	.221**	-0.045
PARTY	0.015	0.021	0.006	0.038	-0.001
Party: Dem	-0.063	0.011	0.025	-.123**	.084*
Party: Rep	.072*	-0.068	0.003	.118**	-0.051
Party: DTS	-0.007	.092**	-0.03	0.033	-0.034
Party: Other	-0.002	0.001	-0.02	-0.014	-0.03
Political Philosophy	-0.043	.077*	0.039	0.048	-0.014

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Education: Grades 9 to 11	Education: HSGrad	Education: Some College	Education: College Grad	Education: Post Grad	Home Owner
EMPLOYMENT STATUS	-0.068	-0.066	-0.053	0.045	.119**	0.025
Employment: Student	-0.037	0.037	.083*	-0.024	-.079*	-.249**
Employment: Homemaker	0.018	-.071*	0.007	0.067	-0.031	0.063
Employment: Retired	.098**	.099**	-0.043	-0.063	-0.046	.147**
Employment: Part Time	-0.049	-0.017	.101**	-0.052	-0.008	-0.006
Employment: Full Time	-.072*	-.075*	-.083*	.085*	.116**	-0.026
Employment: Unemployed	0.068	0.065	0.057	-.071*	-0.062	-0.062
EDUCATION	-.353**	-.580**	-.255**	.347**	.666**	0.048
Education: Grades 1 to 8	-0.017	-0.047	-0.068	-.074*	-0.051	0.032
Education: Grades 9 to 11	1	-.071*	-.101**	-.111**	-.076*	-0.042
Education: HSGrad	-.071*	1	-.282**	-.308**	-.211**	-0.005
Education: Some College	-.101**	-.282**	1	-.439**	-.300**	-0.036
Education: College Grad	-.111**	-.308**	-.439**	1	-.329**	-0.008
Education: Post Grad	-.076*	-.211**	-.300**	-.329**	1	0.068
Home_Owner	-0.042	-0.005	-0.036	-0.008	0.068	1
Yrs in Contra Costa	.088*	.088*	0.068	-.125**	-0.065	.160**
INCOME	-.214**	-.222**	-.111**	.187**	.224**	.369**
Income: 10K or Less	.144**	0.025	.075*	-.089*	-0.055	-.165**
Income: 10,001 - 20,000	0.055	.094**	0.055	-.106**	-.075*	-.143**
Income: 20,001 - 30,000	.178**	0.066	-0.008	-0.041	-.091*	-.127**
Income: 30,001 - 40,000	0.003	0.028	0.021	-0.015	-0.033	-.070*

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Education: Grades 9 to 11	Education: HSGrad	Education: Some College	Education: College Grad	Education: Post Grad	Home Owner
Income: 40,001 - 50,000	0.016	0.029	0.001	0.008	-0.035	-.079*
Income: 50,001 - 60,000	-0.026	0.036	0.062	-0.057	-0.025	-0.015
Income: 60,001 - 70,000	-0.051	0.011	0.001	0.04	-0.027	0.014
Income: 70,001 - 80,000	-0.044	-0.056	0.043	-0.002	0.034	0.038
Income: 80,001 - 100,000	-0.051	-.106**	-0.047	0.067	.096**	.116**
Income: Over 100,000	-0.068	-.141**	-.074*	.102**	.142**	.195**
PARTY	-.084*	-0.029	0.035	0.013	-0.005	-0.049
Party: Dem	.130**	0.059	0	-.114**	0.009	-0.007
Party: Rep	-.115**	-0.048	-0.038	.112**	0.008	.114**
Party: DTS	0.003	-0.016	-0.004	0.006	0.025	-0.052
Party: Other	-0.045	-0.009	.075*	0.008	-0.06	-.135**
Political Philosophy	0.008	-0.004	-0.043	-0.02	.076*	-.113**

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Yrs in Contra Costa	INCOME	Income: 10K or Less	Income: 10,001 - 20,000	Income: 20,001 - 30,000	Income: 30,001 - 40,000
EMPLOYMENT STATUS	-.147**	.185**	-.134**	-.110**	-0.025	-0.018
Employment: Student	-.088*	-.088*	.144**	0.048	-0.062	0.004
Employment: Homemaker	-.123**	.150**	0	-0.028	-0.037	-0.057
Employment: Retired	.378**	-.282**	0.037	.111**	.114**	.079*
Employment: Part Time	-0.055	0.005	0.014	0.012	-0.013	-0.024
Employment: Full Time	-.194**	.209**	-.118**	-.123**	-0.048	-0.015
Employment: Unemployed	0.004	-0.056	0.049	0.047	0.012	-0.03
EDUCATION	-.178**	.397**	-.119**	-.174**	-.172**	-0.042
Education: Grades 1 to 8	.083*	-.086*	-0.02	.106**	0.065	0.006
Education: Grades 9 to 11	.088*	-.214**	.144**	0.055	.178**	0.003
Education: HSGrad	.088*	-.222**	0.025	.094**	0.066	0.028
Education: Some College	0.068	-.111**	.075*	0.055	-0.008	0.021
Education: College Grad	-.125**	.187**	-.089*	-.106**	-0.041	-0.015
Education: Post Grad	-0.065	.224**	-0.055	-.075*	-.091*	-0.033
Home_Owner	.160**	.369**	-.165**	-.143**	-.127**	-.070*
Yrs in Contra Costa	1	-.193**	0.007	0.068	0.065	.090*
INCOME	-.193**	1	-.421**	-.349**	-.379**	-.320**
Income: 10K or Less	0.007	-.421**	1	-0.037	-0.052	-0.062
Income: 10,001 - 20,000	0.068	-.349**	-0.037	1	-0.053	-0.063
Income: 20,001 - 30,000	0.065	-.379**	-0.052	-0.053	1	-.088*
Income: 30,001 - 40,000	.090*	-.320**	-0.062	-0.063	-.088*	1

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Yrs in Contra Costa	INCOME	Income: 10K or Less	Income: 10,001 - 20,000	Income: 20,001 - 30,000	Income: 30,001 - 40,000
Income: 40,001 - 50,000	0.048	-.162**	-0.054	-0.055	-.077*	-.091**
Income: 50,001 - 60,000	0.014	-0.054	-0.063	-0.064	-.089*	-.106**
Income: 60,001 - 70,000	-0.061	.079*	-0.06	-0.061	-.086*	-.103**
Income: 70,001 - 80,000	-0.013	.181**	-0.052	-0.053	-.075*	-.089*
Income: 80,001 - 100,000	-0.057	.341**	-0.06	-0.061	-.086*	-.103**
Income: Over 100,000	-.132**	.634**	-.080*	-.082*	-.115**	-.137**
PARTY	-.105**	0.039	-0.018	-0.005	-0.05	-0.022
Party: Dem	.111**	-.152**	0.03	.089*	.078*	0.064
Party: Rep	0.022	.183**	-0.05	-0.054	-.080*	-.115**
Party: DTS	-.156**	0.024	-0.015	-0.039	-0.02	0.027
Party: Other	-.077*	-0.062	0.05	-0.029	0.018	0.055
Political Philosophy	-0.064	-0.033	-0.038	-0.006	0.066	0.053

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Income: 40,001 - 50,000	Income: 50,001 - 60,000	Income: 60,001 - 70,000	Income: 70,001 - 80,000	Income: 80,001 - 100,000	Income: Over 100,000
EMPLOYMENT STATUS	0.024	0.057	.072*	.097**	0.043	0.06
Employment: Student	0.001	0.002	-0.013	-0.018	-0.013	-0.033
Employment: Homemaker	-0.06	-0.058	-0.037	0	.081*	.114**
Employment: Retired	-0.001	-0.016	-0.065	-.108**	-.096**	-.146**
Employment: Part Time	0.033	-0.026	0.057	0.003	-0.005	0.007
Employment: Full Time	0.001	0.069	0.05	.105**	0.067	.074*
Employment: Unemployed	0.037	-0.031	-0.003	-0.018	-0.055	0.011
EDUCATION	-0.027	-0.048	0.018	0.062	.143**	.221**
Education: Grades 1 to 8	-0.03	0.005	-0.034	-0.029	0.007	-0.045
Education: Grades 9 to 11	0.016	-0.026	-0.051	-0.044	-0.051	-0.068
Education: HSGrad	0.029	0.036	0.011	-0.056	-.106**	-.141**
Education: Some College	0.001	0.062	0.001	0.043	-0.047	-.074*
Education: College Grad	0.008	-0.057	0.04	-0.002	0.067	.102**
Education: Post Grad	-0.035	-0.025	-0.027	0.034	.096**	.142**
Home_Owner	-.079*	-0.015	0.014	0.038	.116**	.195**
Yrs in Contra Costa	0.048	0.014	-0.061	-0.013	-0.057	-.132**
INCOME	-.162**	-0.054	.079*	.181**	.341**	.634**
Income: 10K or Less	-0.054	-0.063	-0.06	-0.052	-0.06	-.080*
Income: 10,001 - 20,000	-0.055	-0.064	-0.061	-0.053	-0.061	-.082*
Income: 20,001 - 30,000	-.077*	-.089*	-.086*	-.075*	-.086*	-.115**
Income: 30,001 - 40,000	-.091**	-.106**	-.103**	-.089*	-.103**	-.137**

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Income: 40,001 - 50,000	Income: 50,001 - 60,000	Income: 60,001 - 70,000	Income: 70,001 - 80,000	Income: 80,001 - 100,000	Income: Over 100,000
Income: 40,001 - 50,000	1	-.093**	-.089*	-.077*	-.089*	-.119**
Income: 50,001 - 60,000	-.093**	1	-.104**	-.090*	-.104**	-.139**
Income: 60,001 - 70,000	-.089*	-.104**	1	-.087*	-.100**	-.134**
Income: 70,001 - 80,000	-.077*	-.090*	-.087*	1	-.087*	-.116**
Income: 80,001 - 100,000	-.089*	-.104**	-.100**	-.087*	1	-.134**
Income: Over 100,000	-.119**	-.139**	-.134**	-.116**	-.134**	1
PARTY	-0.036	-0.028	.072*	-0.031	0.002	-0.013
Party: Dem	0.027	0.038	-0.044	-0.006	0	-.075*
Party: Rep	0	-0.066	0.039	0.031	-0.034	.118**
Party: DTS	-0.042	0.024	0.047	-0.004	.077*	-0.052
Party: Other	-0.006	0.02	-0.04	-0.04	-0.023	-0.012
Political Philosophy	-0.003	0.033	0.02	-.077*	0.007	0.005

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	PARTY	Party: Dem	Party: Rep	Party: DTS	Party: Other	Political Philosophy
EMPLOYMENT STATUS	0.025	-0.002	-0.023	0.06	-0.022	0.041
Employment: Student	0.008	-0.02	-.079*	.084*	.089*	.090*
Employment: Homemaker	-0.025	0.008	0.028	-0.039	-0.022	-0.032
Employment: Retired	-0.024	0.022	0.058	-.110**	-0.023	-.105**
Employment: Part Time	0.015	-0.063	.072*	-0.007	-0.002	-0.043
Employment: Full Time	0.021	0.011	-0.068	.092**	0.001	.077*
Employment: Unemployed	0.006	0.025	0.003	-0.03	-0.02	0.039
EDUCATION	0.038	-.123**	.118**	0.033	-0.014	0.048
Education: Grades 1 to 8	-0.001	.084*	-0.051	-0.034	-0.03	-0.014
Education: Grades 9 to 11	-.084*	.130**	-.115**	0.003	-0.045	0.008
Education: HSGrad	-0.029	0.059	-0.048	-0.016	-0.009	-0.004
Education: Some College	0.035	0	-0.038	-0.004	.075*	-0.043
Education: College Grad	0.013	-.114**	.112**	0.006	0.008	-0.02
Education: Post Grad	-0.005	0.009	0.008	0.025	-0.06	.076*
Home_Owner	-0.049	-0.007	.114**	-0.052	-.135**	-.113**
Yrs in Contra Costa	-.105**	.111**	0.022	-.156**	-.077*	-0.064
INCOME	0.039	-.152**	.183**	0.024	-0.062	-0.033
Income: 10K or Less	-0.018	0.03	-0.05	-0.015	0.05	-0.038
Income: 10,001 - 20,000	-0.005	.089*	-0.054	-0.039	-0.029	-0.006
Income: 20,001 - 30,000	-0.05	.078*	-.080*	-0.02	0.018	0.066
Income: 30,001 - 40,000	-0.022	0.064	-.115**	0.027	0.055	0.053

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	PARTY	Party: Dem	Party: Rep	Party: DTS	Party: Other	Political Philosophy
Income: 40,001 - 50,000	-0.036	0.027	0	-0.042	-0.006	-0.003
Income: 50,001 - 60,000	-0.028	0.038	-0.066	0.024	0.02	0.033
Income: 60,001 - 70,000	.072*	-0.044	0.039	0.047	-0.04	0.02
Income: 70,001 - 80,000	-0.031	-0.006	0.031	-0.004	-0.04	-.077*
Income: 80,001 - 100,000	0.002	0	-0.034	.077*	-0.023	0.007
Income: Over 100,000	-0.013	-.075*	.118**	-0.052	-0.012	0.005
PARTY	1	-.547**	.241**	.303**	.272**	-.190**
Party: Dem	-.547**	1	-.707**	-.318**	-.279**	.368**
Party: Rep	.241**	-.707**	1	-.230**	-.202**	-.433**
Party: DTS	.303**	-.318**	-.230**	1	-.091*	0.045
Party: Other	.272**	-.279**	-.202**	-.091*	1	0.028
Political Philosophy	-.190**	.368**	-.433**	0.045	0.028	1

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	BIRTH YEAR	Age: 18 - 29	Age: 30 - 39	Age: 40 - 49	Age: 50 - 64	Age: 65 and Over
EMPLOYMENT STATUS	-.230**	-.130**	.198**	.252**	0.05	-.415**
Employment: Student	-.300**	.443**	-.086*	-0.065	-.094**	-.109**
Employment: Homemaker	-0.051	-0.005	0.068	0	-0.064	-0.018
Employment: Retired	.639**	-.215**	-.265**	-.272**	0.065	.688**
Employment: Part Time	-.122**	0.055	0.058	0.01	-0.001	-.123**
Employment: Full Time	-.304**	-0.037	.183**	.255**	0.02	-.446**
Employment: Unemployed	-0.041	0.001	0.045	-0.012	-0.001	-0.043
EDUCATION	-0.022	-.107**	0.057	.084*	.077*	-.140**
Education: Grades 1 to 8	.109**	-0.041	-0.053	-0.006	-0.028	.133**
Education: Grades 9 to 11	.099**	-0.014	-0.02	-.092**	-0.027	.151**
Education: HSGrad	0.02	0.04	-0.029	-0.051	0.025	0.041
Education: Some College	-.092**	.145**	-0.04	-0.015	-0.061	-0.015
Education: College Grad	-.077*	-0.052	.093**	.087*	-0.051	-.095**
Education: Post Grad	.109**	-.125**	-0.019	0.006	.132**	-0.008
Home_Owner	.334**	-.338**	-.103**	.094**	.167**	.113**
Yrs in Contra Costa	.448**	-.210**	-.238**	-0.067	.128**	.363**
INCOME	-0.064	-.171**	.081*	.196**	.123**	-.293**
Income: 10K or Less	-0.052	.170**	-0.061	-.077*	-.101**	.085*
Income: 10,001 - 20,000	0.063	0.025	-0.013	-0.064	-.086*	.167**
Income: 20,001 - 30,000	0.054	0.043	-0.037	-0.064	-0.06	.139**
Income: 30,001 - 40,000	0.013	0.066	-0.034	-0.066	0.014	0.055

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	BIRTH YEAR	Age: 18 - 29	Age: 30 - 39	Age: 40 - 49	Age: 50 - 64	Age: 65 and Over
Income: 40,001 - 50,000	-0.006	-0.051	0.062	-0.006	0.024	-0.035
Income: 50,001 - 60,000	-0.043	0.037	-0.027	0.037	0.03	-0.069
Income: 60,001 - 70,000	-.076*	-0.005	.091**	0.01	0	-.083*
Income: 70,001 - 80,000	-.093**	-0.032	.120**	0.037	-0.003	-.118**
Income: 80,001 - 100,000	-0.009	-0.018	-0.028	0.061	.074*	-.083*
Income: Over 100,000	-0.023	-.079*	0.024	.115**	0.013	-.101**
PARTY	-.074*	0.037	0.048	-0.052	-0.021	-0.054
Party: Dem	.073*	-0.051	-.088*	.091**	0.028	0.016
Party: Rep	.099**	-.086*	0.018	-.078*	0.049	.070*
Party: DTS	-.170**	.120**	.075*	0.005	-0.056	-.108**
Party: Other	-.127**	.120**	0.05	-0.039	-.080*	-0.035
Political Philosophy	-.095**	0.06	-0.045	.128**	-0.027	-.113**

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Gender: Female	ETHNICITY	Ethnicity: White/ Caucasian	Ethnicity: Black/ African American	Ethnicity: Hispanic/Latino	Ethnicity: Asian
EMPLOYMENT STATUS	-.102**	0.02	-0.026	0.041	0.006	-0.03
Employment: Student	0.007	0.069	-.097**	-0.018	.100**	.125**
Employment: Homemaker	.249**	-0.052	0.044	0	-0.035	0.016
Employment: Retired	-.100**	-0.054	.089*	-0.039	-0.067	-.072*
Employment: Part Time	.189**	0.025	-0.018	-0.049	0.061	0.003
Employment: Full Time	-.129**	0.023	-0.043	0.056	0.012	0.009
Employment: Unemployed	-0.064	0.018	-0.016	0.041	-0.046	-0.036
EDUCATION	-0.029	-0.063	.152**	-.115**	-.152**	-0.015
Education: Grades 1 to 8	-0.043	0.029	-.098**	.110**	0.067	0.037
Education: Grades 9 to 11	0.021	0.037	-.086*	.144**	-0.01	-0.033
Education: HSGrad	0.004	0.009	-.070*	0.05	.112**	0.026
Education: Some College	0.038	0.003	0.009	-0.065	0.065	-0.048
Education: College Grad	-0.004	0.015	0.028	-0.033	-.088*	.071*
Education: Post Grad	-0.048	-.071*	.094**	-0.016	-.087*	-0.05
Home_Owner	-0.039	-0.032	.099**	-.123**	-0.028	-0.061
Yrs in Contra Costa	-0.042	-.080*	0.057	0.025	0.027	-.136**
INCOME	-.092*	-0.039	.139**	-.179**	-.086*	0.018
Income: 10K or Less	0.069	0.018	-.087*	.161**	0.004	-0.005
Income: 10,001 - 20,000	0.007	0	-0.066	.104**	0.056	-0.007
Income: 20,001 - 30,000	0.036	0.029	-0.039	0.003	0.027	0.018
Income: 30,001 - 40,000	-0.048	-0.015	0	0.011	0.017	-0.024

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

	Gender: Female	ETHNICITY	Ethnicity: White/ Caucasian	Ethnicity: Black/ African American	Ethnicity: Hispanic/Latino	Ethnicity: Asian
Income: 40,001 - 50,000	0.053	-0.009	-0.037	0.035	.079*	0.014
Income: 50,001 - 60,000	0.012	0.008	-0.014	0.025	-0.003	-0.005
Income: 60,001 - 70,000	-0.043	-0.014	0.032	-0.053	-0.032	0.065
Income: 70,001 - 80,000	-0.019	-0.005	-0.014	0.002	0.025	0.017
Income: 80,001 - 100,000	-0.061	0.019	0.012	-0.019	-0.067	0.022
Income: Over 100,000	-0.039	-0.048	.086*	-.088*	-0.028	0
PARTY	-0.022	.086*	0.012	-.088*	-.102**	0.028
Party: Dem	.071*	0.029	-.158**	.209**	.109**	0.009
Party: Rep	-0.062	-0.055	.165**	-.196**	-0.053	-.082*
Party: DTS	-0.017	0.051	-0.022	-0.038	-0.051	.127**
Party: Other	-0.005	-0.012	0.028	-0.002	-0.056	-0.01
Political Philosophy	.071*	-0.004	-0.063	.122**	-0.001	0.051

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Ethnicity: Other	Phone: Unlisted
EMPLOYMENT STATUS	-0.032	0.063
Employment: Student	0.05	0.04
Employment: Homemaker	-0.026	0.005
Employment: Retired	-0.011	-.139**
Employment: Part Time	0.038	0.002
Employment: Full Time	-0.05	.078*
Employment: Unemployed	.089*	0.063
EDUCATION	-0.046	-.097**
Education: Grades 1 to 8	-0.02	-0.039
Education: Grades 9 to 11	0.057	0.01
Education: HSGrad	-0.011	0.068
Education: Some College	.075*	0.035
Education: College Grad	-0.06	0.037
Education: Post Grad	-0.02	-.138**
Home_Owner	-0.006	-.140**
Yrs in Contra Costa	-0.018	-0.057
INCOME	-0.047	-.119**
Income: 10K or Less	0.001	0.036
Income: 10,001 - 20,000	-0.001	0.047
Income: 20,001 - 30,000	.083*	0.022
Income: 30,001 - 40,000	0.031	0.003

Appendix A: Correlation Matrix

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

	Ethnicity: Other	Phone: Unlisted
Income: 40,001 - 50,000	-0.028	.085*
Income: 50,001 - 60,000	-0.017	0.017
Income: 60,001 - 70,000	0.011	0.001
Income: 70,001 - 80,000	0.054	-0.003
Income: 80,001 - 100,000	0.034	-0.057
Income: Over 100,000	-0.061	-.070*
PARTY	0.024	0.044
Party: Dem	0.043	-0.032
Party: Rep	-0.064	-0.052
Party: DTS	0.009	0.044
Party: Other	0.024	.107**
Political Philosophy	-0.008	0.031

APPENDIX B

Text of Survey Questions

INTRO

Hello, may I speak with _____. Hello, my name is _____. I'm calling from _____, an independent public opinion research firm. We're conducting a short research study about issues that concern residents of California, and we would like to include your opinions.

1. I'd like to ask you about some specific recent elections. Did you vote in the November 1996 Presidential election? (IF DON'T RECALL, PROMPT) In that election the candidates for President were Bob Dole and Bill Clinton. With that in mind, do you recall whether or not you voted in the November 1996 Presidential election?

   Voted 1
   Didn't vote 2
   Don't recall 3
   Wasn't registered 4
   Refused 5

2. Did you vote in the November 1992 Presidential election? (IF DON'T RECALL, PROMPT) In that election, the candidates for president were George Bush and Bill Clinton. With that in mind, do you recall whether or not you voted in the November 1992 Presidential election?

   Voted 1
   Didn't vote 2
   Don't recall 3
   Wasn't registered 4
   Refused 5

3. Did you vote in the November 1994 general election for Governor? (IF DON'T RECALL, PROMPT) In that election, the candidates for Governor were Kathleen Brown and Pete Wilson. With that in mind, do you recall whether or not you voted in the November 1996 general election for Governor?

   Voted 1
   Didn't vote 2
   Don't recall 3
   Wasn't registered 4
   Refused 5

4. Have you registered to vote in California since the November 1996 general election, or were you already registered to vote in California before the November 1996 general election?

   Since 1996 1
   Before 1996 2
   DK 3

5. Did you vote in the recent general election for Governor? (IF DON'T RECALL, PROMPT) In that election, the candidates for Governor were Gray Davis and Dan Lungren. With that in mind, do you recall whether or not you voted in the recent general election for Governor?

   Voted 1
   Didn't vote 2
   Don't recall 3
   Wasn't registered 4
   Refused 5

6. Do you feel that things in _____ are going in the right direction, or are they seriously off on the wrong track?

	Right Dir.	Wrong Track	DK
a) Your community	1	2	3
b) California	1	2	3
c) The country	1	2	3

7. Recently, California changed its laws to create what's called an "open" primary election system. Under this new system, voters in a primary election may vote for any candidate of any party, regardless of which party the voter is registered with. (ROTATE)

   [] Supporters of this "open" primary say it increases voter turnout because it allows voters to vote for whoever they really want, without being limited by party registration.

   [] Opponents of the so-called "open" primary say it violates the first amendment freedom of association by forcing voters of one party to allow outsiders to influence

their choice of their party's candidate, even if those outsiders are from an opposing party.

With these arguments in mind, do you support or oppose the "open" primary system for California? (IF SUPPORT/OPPOSE:) Is that strongly, or only somewhat?

Strongly support 1
Somewhat support 2
Somewhat oppose 3
Strongly oppose 4
(DK) 5

8. Currently, the national rules of both the Republican and Democratic parties prohibit the results of "open" primaries from being counted toward selecting party nominees for President. This means that, even though California has moved its primary to an earlier date specifically to increase the state's influence in the presidential selection process, the California primary will be only a "beauty contest" and will no even be counted in selecting party candidates for President.

   After hearing this, would you favor or oppose changing the California primary to allow it to be counted when the national Democratic and Republican parties choose their candidates for president? (IF FAVOR/OPPOSE:) Is that strongly or only somewhat?

   Strongly favor 1
   Somewhat favor 2
   Somewhat oppose 3
   Strongly oppose 4
   (DK) 5

9. Thinking ahead to the election for President in the year 2000, who would you vote for if the dandidates were (ROTATE) []Vice-President Al gore, the Democrat or []Texas Governor George W. Bush, Jr., the Republican?

   Al Gore 1
   (LEAN GORE) 2
   George Bush, Jr. 3
   (LEAN BUSH) 4

10. This past November California voters approved a statewide 9.2 billion dollar school bond initiative. This measure provides funding for education facilities construction, to relieve overcrowding and accommodate growth in student enrollment, and to repair older schools and for wiring and cabling for education technology. Some of

this funding is only available as matching funds. This means that only school districts which pass qualifying local bond measures will have access to this funding.

School districts in your area may or may not be considering local school bond measures, which would quality for these matching funds. If there were such a bond measure in your community would you support or oppose it if it meant a property tax increase of ____ per year? (IF SUPPORT/OPPOSE:) Is that strongly or only somewhat?

(DO NOT ROTATE)	Strong Sup.	Smwt Sup	Smwt Opp	Strong Opp	DK
a) $53	1	2	3	4	5
b) $42	1	2	3	4	5
c) $36	1	2	3	4	5
d) $27	1	2	3	4	5

MY LAST FEW QUESTIONS ARE FOR COMPARISON PURPOSES ONLY.

11. With what ethnic group do you identify yourself: White or Caucasian, Black or African-American, Hispanic or Latino, Asian or of some other ethnic or racial background?

White/Caucasian....1
Black/African American....2
Hispanic/Latino....3
Asian....4
Other (SPECIFY)....5
(DK/NA/REF)....6

12. What is your current employment status? (READ LIST) Are you:

A student....1
A homemaker....2
Retired....3
Unemployed....4
Employed part time....5
Employed full time....6

13. What was the last level of school you completed?

Grades 1-8 1
Grades 9-11 2
High school graduate (12) 3
Some college/vocational 4
College graduate (4 years) 5
Post graduate work 6
(DK/NA/REF) 7

14. How many years have you lived in Contra Costa County? _____

15. Do you own or rent your home?

Own 1
Rent 2
(DK/NA) 3

16. How would you describe yourself politically? Would you say that you are very Conservative, somewhat Conservative, Moderate, somewhat Liberal, or very Liberal?

Very Conservative 1
Somewhat conservative 2
Moderate 3
Somewhat Liberal 4
Very Liberal 5
(Don't Know) 6

17. In what year were you born?

1980 – 1969 (18-29) 1
1968 – 1959 (30-39) 2
1958 – 1948 (40-49) 3
1947 – 1934 (50-64) 4
1933 or earlier (65 or over) 5
(REFUSED/NA) 6

18. I don't need to know the exact amount, but please stop me when I read the category that includes the total income for your household before taxes in 1997, was it:

$10,000 or under	1
$10,001 to $20,000	2
$20,001 to $30,000	3
$30,001 to $40,000	4
$40,001 to $50,000	5
$50,001 to $60,000	6
$60,001 to $70,000	7
$70,001 to $80,000	8
$80,001 to $100,000	9
Over $100,000	0

19. Is your telephone number listed or unlisted?

Listed	1
Unlisted	2
(DK)	3
Refused	4

REFERENCES

Baldassare, M. (2006). *At Issue: California's exclusive electorate*. San Francisco, CA: Public Policy Institute of California.

Baldassare, M., *et. al.* (2010). *PPIC Statewide Survey: Californians and the Environment*. San Francisco, CA: Public Policy Institute of California.

Bell, J. (2006). The Coming Immigration Deal; Congress Will Follow the Polls. *The Weekly Standard*. Jun. 19, 2006, Vol. 11, No. 38. Accessed December 27, 2010 from http://www.weeklystandard.com/Content/Public/Articles/000/000/012/330rcffh.asp.

Belli, R., Traugott, M., Young, M. and McGonagle, K. (1999). Reducing vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring. *Public Opinion Quarterly*, Volume 63, Spring 1999, Accessed June 24, 2006 from http://links.jstor.org/sici=0033-362X%28199921%2963%3A1%3C90%3ARVOISS%3E2.0.CO%3B2-1.

Belli, R., Traugott, M. and Beckman, M. (2001). What Leads to Vote Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies. *Journal of Official Statistics*, Volume 17, Number 4, pp. 479-498.

Braun, K. (1999). Post-experience Advertising Effects on Consumer Memory. *The Journal of Consumer Research*, Vol. 25, No. 4 (Mar. 1999), pp. 319-334.

Connelly, N., and Brown, T. (1994). Effect of Social Desirability Bias and Memory Recall on Reported Contributions to a Wildlife Income Tax Checkoff Program. *Leisure Sciences*, Volume 16.

Crespi, I. (1997). Attitude Measurement, Theory, and Prediction, *Public Opinion Quarterly*, Vol. 41, No. 3 (Autumn, 1977), pp. 285-294, Accessed October 26, 2010 from http://www.jstor.org/stable/2748567.

Darman, D. (2008). Pollsters Wonder How They Got It Wrong on Hillary Victory. New Hampshire Public Radio, Accessed December 31, 2009 from http://www.nhpr.org/node/14799.

Freedman, P. and Goldstein, K. (1996), Building a Probable Electorate From Preelection Polls: A Two-Stage Approach, *Public Opinion Quarterly*, Volume 60, Winter 1996, Accessed March 4, 2006 from http://links.jstor.org/sici=0033-362X%28199624%2960%3A4%3C574%3ABAPEFP%3E2.0.CO%3B2-E.

Freund, J., and Simon, G. (1995). *Statistics: a First Course* (6th ed.). Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Geissinger, S. (2006). Schwarzenegger, Angelides face long, hot summer in contest for governor. *Oakland Tribune*, Sunday June 11, 2006, Accessed June 11, 2006 from http://www.insidebayarea.com/oaklandtribune/ci_3925235.

Keeter, S. (2008). Poll Power. *Wilson Quarterly*, Autumn 2008, Vol. 32 Issue 4, p56-62, Accessed October 25, 2009 from EBSCOhost (Accession Number 34718836).

Kohut, A. (2009). *But What Do the Polls Show? How Public Opinion Surveys Came to Play a Major Role in Policymaking and Politics*. Washington, D.C.: Pew Research Center. Accessed December 19, 2010 from http://pewresearch.org/pubs/1379/polling-history-influence-policymaking-politics.

Lake, C. and Sosin, J. (1998). Public Opinion Polling and the Future of Democracy. *National Civic Review*, Spring98, Vol. 87 Issue 1, p65, Accessed October 25, 2009 from EBSCOhost (Accession Number 646307).

Manheim, J., Rich, R., and Willnat, L. (2002). *Empirical Political Analysis: Research Methods in Political Science*. New York, N.Y. Addison Wesley Longman, Inc.

Parry, H. and Crossley, H. (1950). Validity of Responses to Survey Questions. *Public Opinion Quarterly*, Vol. 14, No. 1 (Spring 1950), pp. 61-80, Accessed December 5, 2010 from http://www.jstor.org/stable/2745899.

Perry, P. (1960). Election Survey Procedures of the Gallup Poll. *Public Opinion Quarterly*, Vol. 24, No. 3 (Autumn, 1960), pp. 531-542, Accessed October 26, 2010 from http://www.jstor.org/stable/2746730.

Perry, P. (1973). A Comparison of the Voting Preferences of Likely Voters and Likely Nonvoters. *Public Opinion Quarterly*, Vol. 37 Issue 1 (Spring 1973), p99, Accessed October 10, 2010 from http://proxy.lib.csus.edu/login?url=http://search.ebscohost.com.proxy.lib.csus.edu/login.aspx?direct=true&db=aph&AN=5415073&site=ehost-live.

Perry, P. (1979). Certain Problems in Election Survey Methodology. *Public Opinion Quarterly*, Vol. 43, No. 3 (Autumn, 1979), pp. 312-325, Accessed October 26, 2010 from http://www.jstor.org/stable/2748227.

Pollock, P. (2009). *An SPSS Companion to Political Analysis* (3rd ed.). Washington, D.C.: CQ Press.

PR Newswire. (2010). *Poll Shows Most Voters Do Not Support DREAM Act and Oppose Using Lame Duck Session to Pass It*. November 29, 2010. Washington, D.C.; Author. Accessed December 27, 2010 from http://www.prnewswire.com/news-releases/poll-shows-most-voters-do-not-support-dream-act-and-oppose-using-lame-duck-session-to-pass-it-111003769.html.

Presser, S. and Traugott, M. (1992). Little White Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting. *Public Opinion Quarterly*, Volume 56, Spring 1992, Accessed December 6, 2010 from http://www.jstor.org/stable/2749222.

Presser, S. (1990). Can Changes in Context Reduce Vote Overreporting in Surveys? *Public Opinion Quarterly*, volume 54, Winter 1990, Accessed March 4, 2006 from http://links.jstor.org/sici=0033-362X%28199024%2954%3a4%3c586%3accicrv%3e2.0.co%3b2-2.

Public Policy Institute of California. (2010). *Just the facts: California's likely voters*. San Francisco, CA: Author.

Schribman, D. (1994). Leadership by the Numbers: Having Brought Polling to New Heights, Will the Clinton Administration Reduce Government to a New Low?

The Boston Globe, Sunday, May 29, 1994, Accessed June 11, 2006 from http://www.pulitzer.org/archives/5649.

Studenmund, A. (2006). *Using Econometrics: A Practical Guide* (5th ed.). Boston, Massachusetts: Pearson Education, Inc.

Traugott, M. and Katosh, J. (1979). Response Validity in Surveys of Voting Behavior, *Public Opinion Quarterly*, Vol. 43, No. 3 (Autumn, 1979), pp. 359-377, Accessed December 5, 2010 from http://www.jstor.org/stable/2745899.

Volgy, T. and Schwarz, J. (1984). Misreporting and Vicarious Political Participation at the Local Level, *Public Opinion Quarterly*, Vol. 48, No. 4 (Winter, 1984), pp. 757-765, Accessed October 29, 2010 from http://www.jstor.org/stable/2748683.