

SOURCES OF VARIATION IN PUBLISHED ELECTION POLLING: A PRIMER

By Cliff Zukin, Professor of Public Policy, Rutgers University
October 2004

The ideas and views contained in this document are those of the author.

Preface

During any campaign season, a multitude of polls about the candidates' relative positions appears in the broadcast and print media. AAPOR representatives are called on repeatedly to explain the how and why of these surveys – especially why their results in the horserace might differ one from another. Cliff Zukin, AAPOR President Elect/Vice President, graciously agreed to prepare the following primer on sources of variations to help us answer inquiries about these polls, and I thank him for his excellent attention to this task. Many thanks are also extended to AAPOR members Larry Hugick, David Moore, Betsy Martin, Michael Traugott, and Jon Krosnick for their considerable assistance and comments in drafting this document. The result of this work is a short primer which goes a long way in explaining election polling to journalists and others. While I am sure some readers will differ with one or another element of this essay -- which is not meant to represent a formal AAPOR position -- it gives journalists and others the background to make sense of what can be perplexing differences in a complex and evolving field.

The primer addresses the surveys conducted by and for the media – not surveys conducted by political pollsters to assist their candidates in conducting winning campaigns. Those campaign-sponsored polls may use many but not precisely the same practices and methods as the published polls described here.

-- Nancy Belden, President, AAPOR

A Primer

Election polling is a special breed among public opinion surveys. It calls for more judgments—the art rather than science of the craft—on the part of the pollster than other types of polls. And this brings into play a host of other reasons why the estimates of well established and well done polls may differ from one another, even when surveys are conducted at a similar point in time.

Real Sampling Error

Most seasoned political observers are familiar with the notion of *sampling error* in public opinion polling. That is, because we select a sample to represent a population we are making an *estimate*, of candidate preference for example, in inferring back to that population. This margin of error, expressed as plus or minus a number of percentage points, is the most commonly known source of variation or why polls may differ. In this election year we often hear statements, such as Bush leads Kerry by five percent, 48% to 43%, with a sampling error of plus or minus four percentage points.

What is less commonly known is that the margin of error does not apply to the *spread* between the two candidates, but to the *percentage point estimates* themselves. If applied to the five point spread the four point margin of error would seem to say that Bush's lead might be as large as nine ($5 + 4$), or as little as one ($5 - 4$). But when correctly applied to the percentage point estimates for the candidates Bush's support could be between 52 and 44% (48 ± 4), and Kerry's between 39 and 47% (43 ± 4). Thus the range between the candidates could be from Bush having a 13 point lead ($52 - 39$) to Kerry having a 3 point advantage ($44 - 47$). So, sampling error is generally much larger than it may seem, and is one of the major reasons why polls may differ, even when conducted around the same time.

Sampling and Coverage Concerns: Household Selection

Sampling is the foundation of survey research, and is based on the branch of mathematics having to do with probability theory. In short, a probability sample is necessary for its numbers to be legitimately generalized from the sample back to the population from which it was drawn. This assumption is not warranted in the case of non-probability (or convenience) samples. Any poll in which respondents were able to select themselves, including call-in polls, and Internet or Web surveys where people volunteer to participate, are non-probability samples, and thus have no basis in science. It is probably a disservice to the public to report the results of such pseudo surveys.

Most pre-election surveys are conducted by telephone, using one of two types of sampling frames, or sources of eligible sampling units (a unit may be a person, a household, *etc.*). The most common approach in the US is what is called a RDD sample, short for random digit dialing. In this case, samples of phone area codes and exchanges are taken, and then random digits added to the end to create 10 digit phone numbers. The first step ensures proper distribution of phone numbers by geography; the second step, adding the random numbers, makes sure that even unlisted numbers are included. This is the standard practiced by almost all public pollsters.

An alternative is called registration based sampling, or RBS. This begins with a sample of individuals drawn from lists of registered voters, to which phone numbers are then matched. This is less costly and more efficient, as almost all calls result in reaching a working phone number and a registered voter, which is not true of an RDD sample. The primary disadvantage of an RBS survey is that it misses people who have recently moved

or have unlisted telephone numbers, which may be a significant portion of the electorate. In New Jersey, for example, an RBS sample would miss approximately 30% of those with landline telephones, who tend to be younger, more urban and more Democratic in their voting behavior. Also, the purging and updating of voter registration lists varies from state to state, so the accuracy of RBS sampling will vary.

Finally, it is important to note that all telephone polls are reaching a somewhat smaller portion of the electorate than in the past because of the growth of cell phones. Since cell phone owners are charged for calls that come into them, these exchanges cannot be included among those available in an RDD sample. A paper presented at the 2004 AAPOR conference estimated that perhaps six percent of households were “cell phone only.” As one might expect, this group is made up disproportionately of younger citizens. Another five percent of households had no cell phone or landline and would not be able to participate in a telephone survey. There is no question that telephone polls miss those people without landline telephones, and we are unsure as to how they may be different from others in terms of their voting behavior. How many in these groups come out on Election Day--and how they are dividing their votes --will be missed by all pre-election telephone polls in 2004. Although weighting may be used to try to make up for this shortfall (see “Weighting” below), “cell phone only” households are a relatively new phenomenon and we do not yet fully understand the consequences of this bias.

Respondent Selection

Sampling a household is only the first stage of RDD surveys, and in itself is insufficient to ensure a representative sample. If interviewers simply spoke with whoever answered the phone the resulting samples would be older and more female than the population as a whole. To produce a representative sample, survey organizations must also go through a respondent selection procedure among potential eligible respondents in the household. Some organizations use a random technique, such as the “last birthday” technique, where the interviewer asks to speak to whoever had the last birthday in the household. There are other techniques of randomization, but the idea is to ensure that everyone has an equal chance for inclusion. Other organizations use a systematic technique, such as asking for the youngest male/oldest female at home, that have produced representative samples in the past.

Some surveys, such as those using automatic or recorded interviewers (touch 1 if you plan to vote for Kerry, 2 if voting for Bush) use *no* selection technique when contacting a household, but instead try to compensate for this by weighting at the end. Like non-probability samples, these types of surveys have little claim of scientific validity and probably should not be reported.

Timing and Field Procedures

Timing of course refers to when a poll is done. As all pollsters are fond of saying, even the most well done poll is a no more than a snapshot in time. Polls do not predict; they describe the situation of the moment. Obviously, polls with different field dates may

yield different results, as voter preferences may change with time. However, a largely invisible reason for differences is that polling organizations have different field procedures. Field procedures refer to the ground rules under which the interviewing is done. And there are some tradeoffs to be made. For example, a field period of seven days would allow for a number of callback attempts to reach the designated respondent before allowing substitution with a new household and respondent. But campaign events may happen in those seven days, making that poll harder to interpret. A three-day poll may focus more narrowly on a particular point in time, but perhaps at the sacrifice of callback attempts. Callbacks matter because one respondent may not be the same as the next; extra field time may be necessary to reach younger voters for example, who may be more Democratic in orientation than others. So factors like the number of callbacks, days of interviewing, and response rates may also be reasons why polls purporting to measure the same thing give different results.

Question Ordering and Wording

It has long been known that the ordering and wording of questions in a survey can affect the results. That is, responses to questions asked early in the interview schedule may affect later ones, as frames of reference are set, and respondents strain to be consistent in their responses to interviewers. For example, a survey that asked respondents a set of questions on the economy before asking for whom they planned to vote could lead to a bias in favor of Kerry, who is stronger on domestic issues. And a line of questioning that asked people about terrorism could lead to a bias in favor of Bush when the “vote” question was finally asked.

In order to minimize this problem, most researchers will ask the horserace question (*If the election were held today, for whom would you vote...*) before any other substantive election question on the survey. This does not include neutral questions about whether people are registered and how interested they are. After all, when people go into the voting booth on November 2 they will have had no warm up questioning. However, perhaps in hopes of simulating the campaign, some polling organizations begin their surveys with substantive policy or election-related questions before asking about vote intentions. When interpreting poll results it is always useful to know the context in which a question was asked. While two polls may have asked the horserace question in the same form, one may have done so after unconsciously pushing some respondents in one direction or the other by earlier questioning. Thus, question ordering also becomes a source of possible variation in the results among published polls.

The wording of questions -- even the horserace question -- may also vary from one poll to the next. Some polls will ask a two-way vote intention question, naming only the major party candidates but recording all answers, while others will explicitly add Ralph Nader's name, or ask about the Green party, or add a response choice of voting for an independent candidate. Some polls ask the horserace question twice in the same survey, once with the two major party candidates and once with a more expansive list of candidates. But one of these must be asked before the other, and the order may influence responses. While most polling organizations asking about the candidates add their party labels as a cue, some

may just name the candidates. And trial heat questions that also name the Vice Presidential candidates may produce somewhat different results than when only Bush and Kerry are mentioned. So differences in question wording also may be a reason why polls have small differences in their reported findings.

Weighting

Weighting is an important and common practice in survey research. Even the best polls cannot interview a perfect sample, due to non-response and non-coverage among other reasons. (Non-response occurs when people refuse to take part in the survey; non-coverage occurs when people cannot be included in the survey – for example, an Internet survey would miss households without computers and a telephone survey would miss those without landlines example.) Thanks to the U.S. Census, we know how many people with fixed characteristics of age, race and education are in the population and we should have in our samples. When we look at whom we actually interview in our samplings, we can adjust -- or weight -- for these characteristics to make sure they are correctly represented.

Weighting likely voters: Because of the Census, we know the characteristics of the general adult population -- but of course, we do not know what the voting public will look like on Election Day. We do not know how many of any racial, educational or generational group will be going to the polls until after the fact. In the 2000 election, the national exit poll estimated that African Americans made up about 10% of the electorate, and they gave about 90% of these votes went to Al Gore. What if seven percent of a polling organization's sample is made up of African Americans in 2004? What if it is 13%? It will obviously make a difference in the horserace estimate, but we will not know which is correct until Election Day. And, of course, the past is no guarantee of the future. So the pollster's dilemma is "What do we weight to?" Most pollsters of published surveys first ask a sample of the general population about their race, age, etc., weight their data to what a random sample of the population should look like, and then go on to pull likely and non-likely voters out of that big (already weighted) general population sample. Some, however, weight to a picture of what they *believe* turnout will be, based on past experience and elections -- and not everyone doing so is painting the same portrait. (This practice is more common among campaign-sponsored polls than published polls.)

Weighting party identification: Another source of differences in pre-election polls is whether the pollster uses party identification as a weighting factor, that is, weighting to reflect an assumed distribution of electorate by the political party of respondents. A party identification question, generally placed near the end of the survey, asks people to state whether they consider themselves a Democrat, Republican, independent or something else. Some pollsters, including this author, do not feel it is appropriate to weight by party. They note that party is not a fixed attribute, like race or gender or age, and that peoples' responses to this question change based on circumstances and events. And indeed, the American public does show fluctuation in

partisanship over time, as well as individual changes. So these pollsters do not consider party ID as a variable for weighting.

Some others weight using an aggregate of their recent phone polls, in an effort to smooth out the ups and downs of party ID that they find in the polls with short field times. Still other pollsters take party ID into consideration either weighting by what they believe party distribution is or weighting their data using party ID estimates from exit polls. This has been a subject of on-going debate among pollsters in recent years and will no doubt continue to be.

Likely Voters

One problem pollsters face here is that not all the respondents who tell us they plan to vote will actually do so. When respondents' self-report of intentions in pre-election polls have been compared to actual turnout (again, known only after the election) we have historically found a large over-report of voting intentions. So the pollsters' dilemma here is to separate the wheat from the chaff: Of all those saying they will vote on Election Day, which ones will really do it, and which ones will stay home? And, of course, people change in their commitment to voting as the campaign unfolds. Respondents are probably better able to tell if they really are going to vote as it gets closer to Election Day. This means that the definition of *likely* voters is somewhat of a moving target, compared to the definition of *registered* voters, for example.

Research finds no magic bullet question or set of questions that can reliably determine likely voters with 100% accuracy. Thus, different organizations have different ways of estimating who are probable voters. Most polls ask a combination of questions that cover self-reported vote intention, measures of engagement (following the election closely, interest) and past behavior (voted in prior elections). They then combine responses to create an index that gives each respondent a total score. Most then use a cutoff point so that only the candidate preferences of the "most" likely voters are used, and the choices of others are discarded. But even while most use such a scale, the component questions that go into the scale differ, and so this too is a source of variation among polling organizations. There are other approaches as well. At least one national poll relies on a single question of reported intention; some may give voters weights based on the probability of voting to all in the sample rather than using a cutoff; some use a fixed set of screening questions that have worked well for them in the past.

A second issue in determining likely voters is estimating *how many* there will be, which may affect the division of the vote. In New Jersey, for example, the percentage of registered voters turning out in the last three presidential elections was 82% in 1992, 72% in 1996, and 70% in 2000. Based on this, a pollster might expect turnout to be about 70% in 2004. But, by comparison, surveys have noted that interest in the election is higher in 2004 than in prior years. So, what should be used as the expected turnout figure -- 70% or 80%, or 85%? Suppose a choice of a cutoff point of 70% gives an estimate that Kerry leads Bush in the state by four percentage points. But when expanding the

expected electorate to 80 %, it may be that the data show Kerry leading by six percentage points. So, another source of possible differences is what percentage of voters is let in during the likely voters scoring process. Moreover, some may start with a base of registered voters (72% in New Jersey in 2000) while others may work with a percentage of the voting age population (52%) as their base. There are also differences in the voting *age* and voting *eligible* populations in each of the states. Thus while all polling organizations will release figures for who they believe are *likely* voters, no two organization will define them in exactly the same way.

In Summary

There are a number of choices to be made in the course of conducting election polling beyond sampling error. We call these “house differences” where different organizations have different ways of doing this type of research. To look for trends it is probably safest to compare polls done by the same organization at different times, rather than to try to compare polls with different methodologies done at different times. Given the unique nature of election polling, it is likely that outsiders may look at them with puzzlement and ask, “What’s going on”? We hope this essay is helpful to our journalistic and other colleagues in understanding some of the sources of variation in election polling. From the inside, those of us conducting election polls see a fair amount of consistency in findings amid the complexity of a science-based-art.

Cliff Zukin may be reached at zukin@rci.rutgers.edu.
Nancy Belden may be reached at nancybelden@brspoll.com.